

# A Rate Optimal Procedure for Sparse Signal Recovery under Dependence \*

Jun Li and Ping-Shou Zhong

Kent State University and Michigan State University

## Abstract

The paper considers the problem of identifying the sparse different components between two high dimensional means of column-wise dependent random vectors. We show that the dependence can be utilized to lower the identification boundary for signal recovery. Moreover, an optimal convergence rate for the marginal false non-discovery rate (mFNR) is established under the dependence. The convergence rate is faster than the optimal rate without dependence. To recover the sparse signal bearing dimensions, we propose a Dependence-Assisted Thresholding and Excising (DATE) procedure, which is shown to be rate optimal for the mFNR with the marginal false discovery rate (mFDR) controlled at a pre-specified level. Simulation studies and case study are given to demonstrate the performance of the proposed signal identification procedure.

**KEYWORDS:** False discovery rate; high dimensional data; multiple testing; thresholding

---

\*Emails: junli@math.kent.edu, pszhong@stt.msu.edu

# 1. INTRODUCTION

In genetic studies, one important task is selecting the differentially expressed genes, which can be crucial in identifying novel biomarkers for cancers. Motivated by the problem of identifying differentially expressed genes, we consider the high dimensional model

$$X_{ij} = \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \stackrel{i.i.d.}{\sim} N(0, \Sigma_i) \text{ for } i = 1, 2 \text{ and } 1 \leq j \leq n_i, \quad (1.1)$$

where  $\mu_i$  is a  $p$  dimensional population mean vector and  $\Sigma_i$  is a  $p \times p$  covariance matrix. If we let  $\delta = \mu_1 - \mu_2 = (\delta_1, \dots, \delta_p)^T$ , our interest is to determine which components of  $\delta$  are non-zero.

Due to high dimensionality and relatively small sample sizes in modern statistical data such as the microarray data, we consider  $p \gg n_i$ . Despite the large number of dimensions, we assume that there are only a small number of signal bearing dimensions. This assumption is thought to be reasonable in many applications. For instance, it is commonly believed that there are only a small number of genes that are significantly differentially expressed between two treatments in a study. Therefore,  $\delta$  is sparse in the sense that most of its components are zero but only a small portion of them are non-zero. In order to recover these sparse signals, a commonly used approach is the multiple testing procedure. Each dimension  $j \in \{1, \dots, p\}$  is tested by a  $t$ -statistic which is expected to have significant value if  $\delta_j \neq 0$  and, conversely, to be insignificant if  $\delta_j = 0$ . After all the p-values associated with the  $t$ -statistics are ranked, the dimensions with p-values smaller than a critical p-value threshold are selected and treated as signal bearing dimensions.

In the multiple testing procedure, the threshold is chosen to control the false discovery rate (FDR), which is defined as the fraction of false positives among all the rejected hypotheses. For this purpose, Benjamini and Hochberg (1995) introduced a novel procedure (BH procedure) which is shown to be more desirable than other

procedures such as the Bonferroni correction that control the familywise error rate (FWER) since the former is less conservative than the latter. However, the BH procedure relies on the assumption that the test statistics corresponding to the true null hypotheses are independent. In real applications, it is also important to consider the effect of dependence on multiple testing. For example, in genetic studies, genes are actually correlated to achieve certain biological tasks due to the internal structure dictated by the genetic networks of living cells (e.g. gene ontology). It has been shown that the presence of the dependence among test statistics can substantially affect the number of reported non-null hypotheses since the empirical null distribution of dependent p-values can be significantly different from the theoretical null distribution under independent assumption (Efron, 2007). Then the outcome of genetic studies by simply ignoring the intergene correlation is implausible, and a clear strategy to control the false positives in the multiple testing for dependent data is needed (Qiu et al., 2005).

Some efforts have been made to address the effect of dependence on the multiple testing by assuming some special dependence structures. For example, Benjamini and Yekutieli (2001) showed that when the test statistics corresponding to the true null hypotheses have the positive regression dependence, the BH procedure asymptotically controls the FDR as well as the independence case. Based on a hidden Markov model for the dependence structure, Sun and Cai (2009) proposed an oracle and an asymptotically optimal data-driven procedures which were shown to be able to minimize the false non-discovery rate (FNR) while controlling the FDR at a pre-specified level. Xie, Cai and Li (2011) established a Bayes oracle rule along with its data adaptive rule based on independent data, which were shown to be optimal in that it minimizes the sum of false negatives and false positives. They also argued that the proposed methods are still valid and remain optimal under short-range dependence.

In this paper, we exploit the nature of dependence differently by investigating its effect on the signal identification boundary without assuming any particular dependence structure. The identification boundary is defined to be a line that separates the plane of signal sparsity and signal strength into two regions. In the region above the line, signals can be recovered individually. But below the line, a successful identification is impossible (Donoho and Jin, 2004, Hall and Jin, 2010, and Ji and Jin, 2012). Although the identification boundary for independent data is well established, we are not aware of any existing results exploring the benefits of data dependence in terms of the identification boundary. Here we show that the signal identification boundary by incorporating data dependence is lower than that without existence of dependence. More precisely, the explicit expression for the identification boundary is established when dependence is present.

When identifying the signals, people are interested in the procedure that minimizes the FNR while the FDR is controlled at a certain level. However, in the setting of data dependence, the optimal convergence rate for the FNR is still largely unknown. Not only does the paper demonstrate the benefits of data dependence in terms of the identification boundary, but it also establishes the optimal convergence rate for the marginal false non-discovery rate (mFNR) under dependence, which is shown to be faster than the rate with independent data.

To identify the sparse signals, we propose a dependence-assisted thresholding and excising (DATE) procedure. The procedure is implemented by first transforming the original data through the matrix

$$\Omega = (\omega_{kl}) = \left( \frac{n_2}{n_1 + n_2} \Sigma_1 + \frac{n_1}{n_1 + n_2} \Sigma_2 \right)^{-1}. \quad (1.2)$$

Then, the null components of the transformed data are removed by conducting a marginal thresholding, which is then followed by an additional step to excise the fake signals by maximizing a penalized MLE. As we will show in Section 4, the proposed

procedure attains not only the signal identification boundary under dependence but also the optimal convergence rate for the mFNR with the marginal false discovery rate (mFDR) controlled at a pre-selected level, and thus is superior compared with other methods without taking data dependence into account.

The rest of the paper is organized as follows. In Section 2, we establish two lower bounds: one for the risk function (2.2) and another for the convergence rate of the mFNR. To show the optimality of these two bounds, we first demonstrate the benefit of transforming the data by the matrix  $\Omega$  in (1.2) in Section 3. Then a thresholding and excising procedure based on the transformed data is introduced in Section 4. The proposed procedure is shown to be able to achieve two lower bounds established in Section 2 and thus is rate optimal. Section 5 illustrates some numerical studies and Section 6 reports an empirical study to select differentially expressed genes for a human breast cancer data set. Discussion is given in Section 7. All technical details are relegated to the Appendix.

## 2. LOWER BOUNDS FOR SIGNAL IDENTIFICATION UNDER DEPENDENCE

To establish the lower bound of the signal identification boundary in the dependent setting, we start with some notations and definitions. Denote  $S_\beta = \{k : \delta_k \neq 0\}$  to be a set including the locations of the non-zero  $\delta_k$ . The number of non-zero elements in  $S_\beta$  is  $p^{1-\beta}$  for  $\beta \in (\frac{1}{2}, 1)$ . Define  $L_p$  to be a slowly varying logarithmic function in the form of  $(a \log p)^b$ . Without loss of generality, we assume both  $\Sigma_1$  and  $\Sigma_2$  are standardized to have unit diagonal elements. With matrix  $\Omega = (\omega_{ij})$  defined in (1.2), let

$$\underline{\omega} = \lim_{p \rightarrow \infty} \min_{1 \leq k \leq p} \omega_{kk}, \quad \text{and} \quad \bar{\omega} = \overline{\lim}_{p \rightarrow \infty} \max_{1 \leq k \leq p} \omega_{kk}. \quad (2.1)$$

We model  $\delta$  to satisfy the following condition (see Ji and Jin, 2012):

(C1). The components of  $\delta$  follow a mixture distribution

$$\delta_k \stackrel{i.i.d.}{\sim} (1 - p^{-\beta})h_0 + p^{-\beta}\pi_p, \quad k = 1, \dots, p,$$

where  $h_0$  is a point mass at 0 and  $\pi_p$  is a distribution with the support  $[-\sqrt{2r\log p/n}, 0) \cup (0, \sqrt{2r\log p/n}]$  for  $r > 0$  and  $n = \frac{n_1 n_2}{n_1 + n_2}$ .

In the independent case, the identification boundary that describes the relationship between signal sparsity  $\beta$  and signal strength  $r$  is defined to be a line  $r = \beta$  in the  $\beta$ - $r$  plane. In the region above the line, it is possible to identify them individually, but it becomes impossible in the region below the line. Since stronger magnitude of signals is needed to discover non-zero components individually, the identification boundary lies above the detection boundary that separates the  $\beta$ - $r$  plane into the so-called detectable region and undetectable region.

Given  $\delta_k$  for  $1 \leq k \leq p$ ,  $\hat{\delta}_k$  is denoted as an estimate of  $\delta_k$ . For any signal identification procedure, there are generally two types of error related with the signal estimate  $\hat{\delta}_k$ : the false negative meaning that  $\delta_k \neq 0$  but  $\hat{\delta}_k = 0$ , and the false positive representing that  $\delta_k = 0$  but  $\hat{\delta}_k \neq 0$ . Then the optimal procedure for signal recovery can be defined as the one that minimizes the expected weighted sum of false negatives and false positives:

$$H(\Lambda) = \mathbb{E} \left\{ \sum_{k \in S_\beta} \mathbb{I}(\hat{\delta}_k = 0) + p^{-\Lambda} \sum_{l \in S_\beta^c} \mathbb{I}(\hat{\delta}_l \neq 0) \right\}, \quad (2.2)$$

where the weight  $p^{-\Lambda}$  with  $\Lambda \in [0, \infty)$  is chosen to adjust the level of false positives. If  $\Lambda = 0$ , there is no preference on either the false positives or the false negatives, and the risk (2.2) becomes the misclassification error adopted by Ji and Jin (2012) for establishing the optimal convergence rate for the variable selection problem in the high-dimensional regression model. On the other hand, choosing a larger value of  $\Lambda$  leads to a smaller weight function  $p^{-\Lambda}$ , which consequently allows the optimal

procedure to produce relatively larger false positives when minimizing  $H(\Lambda)$ . The effect of  $\Lambda$  on false positives can be demonstrated by Figure 1. Assume that the minimization of  $H(0)$  is achieved at the intersection point *diamond* of the false positives line (FP) and the false negatives line (FN). By multiplying FP with  $p^{-\Lambda}$  (dash line), the FP becomes less important in  $H(\Lambda)$  and  $H(\Lambda)$  is minimized at the intersection point *star* which is on the right side of the intersection point *diamond*. As a result, the expected false positives corresponding to the minimized  $H(\Lambda)$  is larger than that corresponding to the minimized  $H(0)$ . The universal lower bound of the risk function  $H(\Lambda)$  at a fixed value  $\Lambda$  is established by the following theorem.

**Theorem 1.** Assume condition (C1) and the model (1.1) for  $X_{ij}$ . As  $p \rightarrow \infty$ ,

$$H(\Lambda) \geq \begin{cases} L_p p^{1-\beta-(\bar{\omega}r-\beta+\Lambda)^2/(4\bar{\omega}r)}, & -r < (\Lambda - \beta)/\underline{\omega} < r \\ p^{1-\beta}, & r < (\beta - \Lambda)/\bar{\omega} \\ p^{1-\Lambda}, & r < (\Lambda - \beta)/\bar{\omega} \end{cases}$$

where  $\underline{\omega}$  and  $\bar{\omega}$  are defined in (2.1), and  $L_p$  is a slowly varying logarithmic function.

The universal lower bound varies with different values of  $r$ ,  $\beta$  for each fixed value of  $\Lambda$ . If we choose  $\Lambda = 0$ , the misclassification error has the lower bound

$$H(0) \geq \begin{cases} L_p p^{1-\beta-(\bar{\omega}r-\beta)^2/(4\bar{\omega}r)}, & r > \beta/\underline{\omega} \\ p^{1-\beta}. & r < \beta/\bar{\omega} \end{cases}$$

Some key observations are as follows. First, if the signal strength  $r < \beta/\bar{\omega}$ , the misclassification error is no less than  $p^{1-\beta}$ , the number of non-zero  $\delta_k$ , which implies that there exists no successful signal identification procedure. The area  $r < \beta/\bar{\omega}$  in  $r - \beta$  plane is thereafter called the region of no recovery. On the other hand, if the signal strength attains  $r \geq (1 + \sqrt{1 - \beta})^2/\underline{\omega}$ , the misclassification error asymptotically converges to zero and all the signals can be successfully recovered. The corresponding region is called the region of full recovery. The area sandwiched between the no recovery region and the full recovery region satisfies  $\beta/\bar{\omega} < r < (1 + \sqrt{1 - \beta})^2/\underline{\omega}$ ,

having the misclassification error less than the number of signals and greater than zero. This region is called region of partial recovery. Most importantly, since  $\bar{\omega} \geq \underline{\omega} > 1$  under data dependency shown by Lemma 1 in Appendix, the partial recovery boundary  $r = \beta/\bar{\omega}$  and full recovery boundary  $r = (1 + \sqrt{1 - \beta})^2/\underline{\omega}$  used to separate three regions are lower than those without existence of data dependence.

To demonstrate the observations above, we consider  $\Sigma_1 = \Sigma_2 = (\rho^{|i-j|})$  for  $1 \leq i, j \leq p$  in model (1.1) such that the data dependence is exhibited by the value of  $\rho$ . If  $\rho = 0$ ,  $\bar{\omega} = \underline{\omega} = 1$  since there is no data dependence. On the other hand, if  $\rho = 0.6$ , we obtain  $\underline{\omega} = 1.5625$  and  $\bar{\omega} = 2.125$ . The corresponding phase diagrams with and without data dependence are displayed in Figure 2 in which the partial signal identification boundary and the full recovery boundary with  $\rho = 0.6$  are lower than those with  $\rho = 0$  due to the fact that  $\underline{\omega} > 1$  and  $\bar{\omega} > 1$ . As a result, even though the signals with  $r < \beta$  are unable to be identified by any procedure if there exists no data dependence, some of them can be recovered as long as the signal strength  $r > \beta/2.125$  with the existence of data dependence. The benefit to the full signal identification with the existence of dependence can be seen based on the similar derivation.

There is a close connection between the signal recovery and the weighted risk function  $H(\Lambda)$ . It has been shown that by properly choosing  $\Lambda$ , the decision rule that minimizes the weighted risk function  $H(\Lambda)$  is also the optimal procedure that controls the marginal FDR at level  $\alpha$  and minimizes the marginal FNR (mFNR) in the multiple testing problem (Sun and Cai, 2007, Sun and Cai, 2009, and Xie, Cai, Maris and Li, 2011). Let FP= false positives, TP=true positives, FN= false negatives and TN= true negatives. The mFDR and mFNR are defined as

$$\text{mFDR} = \left\{ \frac{E(\text{FP})}{E(\text{FP}) + E(\text{TP})} \right\} \quad \text{and} \quad \text{mFNR} = \left\{ \frac{E(\text{FN})}{E(\text{FN}) + E(\text{TN})} \right\}.$$

Genovese and Wasserman (2002) showed that mFDR and mFNR are asymptotically equivalent to FDR and FNR under weak conditions. In general, the connection



between  $\Lambda$  and  $\alpha$  is complicated. The following theorem provides a solution for choosing a proper  $\Lambda(\alpha)$  such that the mFDR is controlled at the level of  $\alpha < 1$ . Moreover, it establishes a lower bound for the mFNR subject to the constraint that  $\text{mFDR} \leq \alpha$ .

**Theorem 2.** Assume condition (C1) and (1.1) for  $X_{ij}$ . If we choose

$$\Lambda(\alpha) = \underline{\omega}r + \beta - 2\sqrt{\underline{\omega}r\beta\left(1 - \frac{g(\alpha, p)}{\beta}\right)}, \quad \text{where } g(\alpha, p) = \frac{\log\left\{\frac{\alpha}{(1-\alpha)}\sqrt{4\pi\beta\log p}\right\}}{\log p},$$

then as  $p \rightarrow \infty$ ,

$$\text{mFNR} \geq L_p p^{-\beta - \left\{\sqrt{\bar{\omega}r} - \sqrt{\beta - g(\alpha, p)}\right\}^2} \quad \text{and} \quad \text{mFDR} \leq \alpha < 1.$$

Similar to the weighted risk function, the lower bound for the mFDR is accelerated with existence of dependence since  $\bar{\omega} > 1$ . In order to show that the lower bounds in Theorems 1 and 2 are tight, we need to search for a signal identification procedure that is able to attain the universal lower bounds. As we will see in next section, the key for this procedure is to take the data dependence into account, which can be done by transforming the data via the matrix  $\Omega$  defined in (1.2).

### 3. DATA TRANSFORMATION

Some additional assumptions are needed to establish the theoretical performance of the procedure we will introduce in this and next sections.

(C2). The eigenvalues of  $\Sigma_i$  for  $i = 1, 2$  satisfy  $C_0^{-1} \leq \lambda_{\min}(\Sigma_i) \leq \lambda_{\max}(\Sigma_i) \leq C_0$  for some constant  $C_0 > 0$ .

(C3). The matrix  $\Omega$  in (1.2) is presumably sparse and belongs to the class

$$\mathcal{V}(c_p, M_p) = \left\{ \Omega : \|\Omega\|_{L_1} \leq M, \max_{1 \leq j \leq p} \sum_{i=1}^p |\omega_{ij}|^\zeta \leq c \quad \text{for } 0 < \zeta < 1 \right\},$$

where  $M$  and  $c$  are fixed constants.

(C4). As  $n \rightarrow \infty$ ,  $p \rightarrow \infty$  and  $\log p = cn^\theta$  for  $c > 0$  and  $\theta < \frac{1-\zeta}{2-\zeta}$  where  $\zeta$  is defined in (C3).

Conditions (C2) and (C3) define a class of matrices with sparse structures, which is originally proposed by Bickel and Levina (2008b). Condition (C4) specifies the exponential growth of dimension  $p$  with  $n$ . All of these conditions are commonly assumed in the literature.

For signal identification, we need to define a statistic to estimate the magnitude of the signal. To this end, we let  $\bar{X}_i^{(k)} = \sum_{j=1}^{n_i} X_{ij}^{(k)} / n_i$  for  $i = 1, 2$  where  $X_{ij}^{(k)}$  is the  $k$ th component of  $X_{ij}$ . Then a measure of  $n\delta_k^2$  is defined by

$$L_k = n\{\bar{X}_1^{(k)} - \bar{X}_2^{(k)}\}^2, \quad k = 1, \dots, p. \quad (3.1)$$

Since the marginal variances of  $L_k$  are the same, the probability of the non-null component being identified depends on the value of  $L_k$  or essentially its signal strength  $\delta_k$ . The magnitude of  $\delta_k$  can be enhanced by transforming  $X_{ij}$  into  $Z_{ij} = \Omega X_{ij}$  where  $\Omega = (\omega_{kl})$  for  $1 \leq k, l \leq p$  is defined in (1.2). The similar transformation was also considered in Hall and Jin (2010) for their innovated higher criticism test, and Cai, Liu and Xia (2014) for testing the equality of two sample mean vectors.

To appreciate signal enhancement induced by the transformation, we let  $\delta_\Omega$  be the difference in two population mean vectors after the transformation. Then the following relationship holds between  $\delta_\Omega$  and the original signal strength  $\delta$ :

$$\delta_{\Omega,k} = \omega_{kk}\delta_k + \sum_{l \neq k \in S_\beta} \omega_{kl}\delta_l, \quad \text{for } k = 1, \dots, p.$$

Lemma 2 in Appendix shows that for sparse signals and sparse  $\Omega$  assumed in (C3),  $\delta_{\Omega,k} = \omega_{kk}\delta_k + o(n^{-1/2})$ , which implies that if  $k \in S_\beta$ ,

$$\frac{\delta_{\Omega,k}}{\sqrt{\omega_{kk}}} = \sqrt{\omega_{kk}}\delta_k + o(n^{-1/2}). \quad (3.2)$$

This, together with  $\omega_{kk} \geq 1$  in Lemma 1, leads to

$$\frac{\delta_{\Omega,k}}{\sqrt{\omega_{kk}}} \geq \delta_k.$$

Therefore, for signal identification, we propose the following test statistic

$$T_k = \frac{n\{\bar{Z}_1^{(k)} - \bar{Z}_2^{(k)}\}^2}{\omega_{kk}}, \quad k = 1, \dots, p,$$

which is constructed based on the transformed data and has the standardized signal strength  $\sqrt{n}\delta_{\Omega,k}/\sqrt{\omega_{kk}}$  greater than standardized signal strength  $\sqrt{n}\delta_k$  of the test statistic (3.1).

In real applications,  $\Omega$  is unknown and needs to be estimated. When  $\Sigma_1$  and  $\Sigma_2$  are bandable,  $\Omega$  can be estimated through the Cholesky decomposition proposed by Bickel and Levina (2008a). Yuan and Lin (2007) considered an  $L_1$  penalized normal likelihood estimator for the sparse precision matrix. More can be found in Friedman, Hastie and Tibshirani (2008). Cai, Liu and Luo (2011) introduced an CLIME estimator based on the constrained  $L_1$  minimization approach for precision matrix estimate. With estimated  $\hat{\Omega}$ , the transformed signal for  $k \in S_\beta$  is  $\hat{\delta}_{\Omega,k} = \sum_{l \in S_\beta} \hat{\omega}_{kl} \delta_l$ . Similar to  $\delta_{\Omega,k}$  when  $\Omega$  is known, Lemmas 1 and 2 show that under some mild conditions, with probability equal to 1,

$$\frac{\hat{\delta}_{\Omega,k}}{\sqrt{\hat{\omega}_{kk}}} \geq \delta_k,$$

Therefore, we consider the following test statistics based on the transformed data  $\hat{Z}_{ij} = \hat{\Omega} X_{ij}$  as the starting point of the proposed signal identification procedure:

$$\hat{T}_k = \frac{n\{\hat{\bar{Z}}_1^{(k)} - \hat{\bar{Z}}_2^{(k)}\}^2}{\hat{\omega}_{kk}}, \quad k = 1, \dots, p. \quad (3.3)$$

The advantage of the statistics in (3.3) relative to (3.1) is that the standardized signal strength has been enhanced by incorporating the dependence, which potentially increases the probability of weak signals being identified by the signal recovery procedure. However, since  $\delta_{\Omega,k} = \sum_{l \in S_\beta} \omega_{kl} \delta_l$ , a side effect of the transformation is that it generates some fake signals, i.e.,  $\delta_k = 0$  but  $\delta_{\Omega,k} \neq 0$  if  $\omega_{kl} \neq 0$  for some  $l \in S_\beta$ . Therefore, a successful signal recovery procedure benefited by data transformation

requires to remove these fake signals. As we will discuss in next section, fake signals can be successfully excised by a penalized method with  $L_0$  penalty. As revealed by Ji and Jin (2012), this approach is very effective in cleaning fake signals but suffers the computational intensity if dimension  $p$  is large. To reduce the complexity of the original signal selection problem, we first need a dimension reduction procedure, which is fulfilled by a thresholding step as we will discuss in next section.

#### 4. DATE PROCEDURE TO RECOVER SIGNALS

To introduce our signal identification procedure, we first focus on most interesting case where  $\underline{\omega}r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$ . According to Theorem 1, this case indicates that the weighted risk  $H(\Lambda)$  does not converge to zero but is less than  $p^{1-\beta}$ . The corresponding region on  $r-\beta$  plane is the partial recovery under a fixed value  $\Lambda$ . The case  $\underline{\omega}r \geq (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$  corresponding to the full recovery region is an easier problem due to the relatively larger signal strength. We will discuss it at the end of this section.

As we have discussed in the previous section, after data transforming,  $p$  coordinates consist of the signals, fake signals and noise. As the first step of the proposed method for signal recovery, a thresholding is conducted to remove the noise. After all the  $p$  dimensions are checked by a threshold function  $2s\log p$ , we set  $\hat{\delta}_k = 0$  for  $k \in \{1, \dots, p\}$  if and only if

$$\hat{T}_k < 2s\log p, \quad (4.1)$$

where  $s > 0$  is chosen to control the level of the threshold, and the decision on other coordinates with  $\hat{T}_k \geq 2s\log p$  will be made in another step following the thresholding step. Although imposing the threshold is to prevent noise, it can potentially screen out signals and thus produce the false negatives. The following Lemma establishes the upper bound of the expected false negatives generated in the thresholding step

(4.1).

**Lemma 3.** Assume conditions (C1), (C3) and (C4). Let  $s \in (0, \frac{(\omega r + \beta - \Lambda)^2}{4\omega r})$  and  $\beta - \Lambda < \omega r < (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$ . As  $p \rightarrow \infty$ ,

$$\mathbb{E} \left\{ \sum_{k=1}^p (\hat{\delta}_k = 0, \delta_k \neq 0) \right\} \leq L_p p^{1 - \beta - (\omega r - \beta + \Lambda)^2 / (4\omega r)}.$$

Since the error above is no more than the error rate established in Theorem 1 provided that  $\omega = \bar{\omega}$ , it does not affect the rate optimality of the whole identification procedure as long as the error made in the following excising step is under control.

The fake signals generated by the transformation are able to survive from the thresholding if

$$\hat{T}_k \geq 2s \log p, \quad k \notin S_\beta.$$

To excise these fake signals, we implement an  $L_0$  penalization approach, which is originally designed for the regression problem. For the purpose of variable selection, this approach directly penalizes the number of non-zero parameters but is hampered by high dimensionality since it requires an exclusive search of all  $2^p$  submodels and is computationally intensive. However, as we will show in the following, this NP hard problem can be circumvented thanks to an important consequence of conducting the thresholding. To see it, we let  $\mathcal{U}(s)$  be a set including all components survived from the thresholding

$$\mathcal{U}(s) = \{k : \hat{T}_k \geq 2s \log p, 1 \leq k \leq p\}. \quad (4.2)$$

We define  $V_0 = \{1, \dots, p\}$  to be a set of nodes and

$$\Omega^*(i, j) = \hat{\Omega}(i, j) \mathbb{I}_{\{|\hat{\Omega}(i, j)| \geq \log^{-1} p\}} \quad (4.3)$$

to be regularized  $\hat{\Omega}$ . Then according to the Gaussian graph theory, given the precision matrix  $\Omega^*$ , any  $i \neq j \in V_0$  are connected if and only if  $\Omega^*(i, j) \neq 0$ . The following Lemma 4 summarizes the consequence after conducting the thresholding.

**Lemma 4.** Assume the conditions (C1)-(C4). Let  $s \in (0, \frac{(\omega r + \beta - \Lambda)^2}{4\omega r})$  and  $\beta - \Lambda < \underline{\omega} r < (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$ . With probability  $1 - L_p p^{-\beta - (\underline{\omega} r - \beta + \Lambda)^2 / (4\omega r)}$ ,  $\mathcal{U}(s)$  are split into disconnected clusters of size no more than a positive integer  $K$  with respect to  $(V_0, \Omega^*)$ .

According to Lemma 4, the  $L_0$  penalization approach can be effectively adopted to each of self-connected subsets with relatively small size. Let  $I_0 = \{i_1, \dots, i_m\}$  be one of the self-connected subsets with size  $m \leq K$ , and  $\hat{A} = \hat{\Omega}^{I_0, I_0}$  be an  $m \times m$  matrix with  $\hat{\Omega}^{I_0, I_0}(k, l) = \hat{\Omega}(i_k, i_l)$ . To excise the fake signals in  $I_0$ , we find an  $m$ -dimensional vector  $\hat{\delta}(I_0)$  each component of which is equal to either 0 or  $\delta^{date}$  or  $-\delta^{date}$  to minimize the following function:

$$n \left\{ (\bar{\tilde{Z}}_1 - \bar{\tilde{Z}}_2)^{I_0} - \hat{A} \delta \right\}' \hat{A}^{-1} \left\{ (\bar{\tilde{Z}}_1 - \bar{\tilde{Z}}_2)^{I_0} - \hat{A} \delta \right\} + (\lambda^{date})^2 \|\delta\|_0, \quad (4.4)$$

where  $\lambda^{date}$  and  $\delta^{date}$  are two tuning parameters.

After we apply the  $L_0$  penalization approach to all the self-connected subsets, each of  $\delta_k$  for  $k = 1, \dots, p$  is eventually determined by the proposed DATE procedure which can be summarized by the following algorithm.

- (1). Transform data  $X_{ij}$  to obtain  $\hat{Z}_{ij} = \hat{\Omega} X_{ij}$  where  $\hat{\Omega}$  is estimated  $\Omega$ ;
- (2). Conduct the thresholding described by (4.1) such that the coordinates  $k = 1, \dots, p$  are assigned to either  $\mathcal{U}(s)$  or its complement  $\mathcal{U}^c(s)$  where  $\mathcal{U}(s)$  is defined in (4.2). For all  $k \in \mathcal{U}^c(s)$ , we set  $\hat{\delta}_k = 0$ ;
- (3). Allocate  $l \in \mathcal{U}(s)$  into different self-connected subsets  $\{I_0^{(1)}, I_0^{(2)}, \dots, I_0^{(h)}\}$  with respect to  $(V_0, \Omega^*)$ . For  $I_0^{(1)}$ ,  $\delta(I_0^{(1)})$  is equal to  $\hat{\delta}(I_0^{(1)})$  each component of which is chosen to be either 0 or  $\delta^{date}$  or  $-\delta^{date}$  in order to minimize the penalized function (4.4). Repeat the same procedure to other  $I_0^{(j)}$  where  $j \in \{2, \dots, h\}$  to determine  $\delta_l$  for  $l \in \mathcal{U}(s)$ .

To easily measure the performance of the proposed DATE procedure, we further assume the following condition which is analogous to (C1) but requires a slightly stronger signal strength than (C1). A similar strategy was also taken in Ji and Jin (2012) to measure the performance of an UPS procedure for variable selection in the high dimensional regression problem.

(C1)'. Similar to (C1), the components of  $\delta$  follow the mixture distribution with  $\pi_p$  being a distribution on the support  $[-(1 + \eta)\sqrt{2r\log p/n}, -\sqrt{2r\log p/n}] \cup [\sqrt{2r\log p/n}, (1 + \eta)\sqrt{2r\log p/n}]$  where  $\eta \leq \frac{\beta - \Lambda}{\sqrt{C_0 r}} \frac{\sqrt{\beta r}}{\sqrt{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta}}$  and the constant  $C_0$  is defined in (C2).

The following theorem establishes the upper bound of the risk (2.2) for the proposed DATE procedure.

**Theorem 3.** Assume conditions (C2)-(C4) and (C1)'. Choose  $s \in (0, \frac{(\omega r + \beta - \Lambda)^2}{4\omega r})$  and  $\beta - \Lambda < \omega r < (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$ , and set the tuning parameters in (4.4) to be

$$\lambda^{date} = \sqrt{2(\beta - \Lambda)\log p}, \quad \delta^{date} = \sqrt{2r\log p/n}.$$

As  $p \rightarrow \infty$ , the weighted risk (2.2) for the DATE satisfies

$$H(\Lambda) \leq L_p p^{1 - \beta - (\omega r - \beta + \Lambda)^2 / (4\omega r)}.$$

Since  $(\omega r - \beta + \Lambda)^2 / (4\omega r) \leq (\bar{\omega} r - \beta + \Lambda)^2 / (4\bar{\omega} r)$ , the lower bound in Theorem 1 is no greater than the upper bound in Theorem 3. Specially, these two bounds match each other if  $\bar{\omega} = \omega$ , which implies both bounds are tight and thus the DATE procedure is rate optimal in terms of the risk (2.2).

Our ultimate goal is to apply the DATE procedure to signal identification. So we need to ensure that it can successfully control the FDR at any desired level  $\alpha < 1$ . By carefully reviewing the whole procedure, we see that the thresholding step (4.1) is designated to control the false negatives and the success of the FDR control is determined only by the excising step (4.4) where the role is played by the tuning

parameter  $\lambda^{date}$ . Due to the adoption of  $L_0$  penalty, smaller value of  $\lambda^{date}$  allows more toleration for the false positives and thus leads to greater FDR. It turns out that if we subtract an additional term from the  $\lambda^{date}$  in Theorem 3, the mFDR can be successfully controlled at  $\alpha < 1$  and the rate of the mFNR is accordingly established by Theorem 4.

**Theorem 4.** Assume conditions (C2)-(C4) and (C1)'. Choose  $s \in (0, \beta)$ ,  $\beta - \Lambda < \underline{\omega}r < (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$  and  $\Lambda = (\sqrt{\underline{\omega}r} - \sqrt{\beta})^2$ . As  $p \rightarrow \infty$ , by setting the tuning parameters of the DATE as

$$\lambda^{date} = \sqrt{2(\beta - \Lambda)\log p - \Upsilon}, \quad \delta^{date} = \sqrt{2r\log p/n},$$

where

$$\Upsilon = \frac{4\underline{\omega}r}{\underline{\omega}r + \beta - \Lambda} \left( \frac{1}{2} \log \log p + \log \left\{ \frac{\alpha \sqrt{\pi} (\underline{\omega}r + \beta - \Lambda)}{2\sqrt{\underline{\omega}r}(1 - \alpha)} \right\} \right).$$

Then,

$$\text{mFDR} \leq \alpha \quad \text{and} \quad \text{mFNR} \leq L_p p^{-\beta - (\sqrt{\underline{\omega}r} - \sqrt{\beta})^2}.$$

Since  $\bar{\omega}r \geq \underline{\omega}r > \beta$ , the optimal rate of the mFNR in Theorem 2 is not faster than the rate in Theorem 4 and two rates are equal to each other asymptotically if  $\bar{\omega} = \underline{\omega}$ . This, combining with the fact that  $\text{mFDR} \leq \alpha < 1$ , shows that the proposed DATE procedure is optimal in that it minimizes the mFNR subject to the constraint that mFDR is controlled at the desired level  $\alpha < 1$ .

There are three tuning parameters needed to estimated in the proposed signal identification procedure: the level of threshold  $s$  in (4.1), two tuning parameters  $\delta^{data}$  and  $\lambda^{date}$  in (4.4). To select tuning parameters  $\lambda^{data}$  and  $\delta^{date}$ , we estimate the sparsity  $\beta$ , the signal magnitude  $r$  and  $\underline{\omega}$  by the following estimators:

$$\hat{\beta} = -\frac{\log \left\{ \frac{1}{p} \sum_{k=1}^p \mathbf{I}(\hat{T}_k > 2q\log p) \right\}}{\log p}, \quad \hat{r} = \frac{1}{2p^{1-\hat{\beta}} \log p} \sum_{k=1}^p \frac{\hat{T}_k - 1}{\hat{\omega}_{kk}} \mathbf{I}(\hat{T}_k > 2q\log p),$$



and

$$\underline{\hat{\omega}} = \min_{1 \leq k \leq p} \hat{\omega}_{kk}, \quad (4.5)$$

where  $q$  is another threshold level controlling the accuracy of estimate in  $\beta$  and  $r$  and the question of properly choosing both  $s$  and  $q$  is addressed in Theorem 5. With two tuning parameters  $\lambda^{data}$  and  $\mu^{date}$  estimated by plugging the  $\hat{\beta}, \hat{r}, \underline{\hat{\omega}}$  into the expressions defined in Theorem 4, the following theorem shows that the performance of the DATE procedure with estimated parameters (4.5) is asymptotically equivalent to the DATE in Theorem 4.

**Theorem 5.** Assume conditions (C2)-(C4) and (C1)'. As  $p \rightarrow \infty$ , by setting  $s \in (0, \beta)$  in (4.1),  $q \in (\beta, \underline{\omega}r)$  in (4.5) and estimating the tuning parameters as

$$\hat{\lambda} = 2\hat{s}\log p, \quad \hat{\lambda}^{date} = \sqrt{2(\hat{\beta} - \hat{\Lambda})\log p - \hat{\Upsilon}}, \quad \hat{\delta}^{date} = \sqrt{2\hat{r}\log p/n},$$

where

$$\begin{aligned} \hat{\Lambda} &= (\sqrt{\underline{\hat{\omega}}\hat{r}} - \sqrt{\hat{\beta}})^2, \\ \hat{\Upsilon} &= \frac{4\underline{\hat{\omega}}\hat{r}}{\underline{\hat{\omega}}\hat{r} + \hat{\beta} - \hat{\Lambda}} \left( \frac{1}{2} \log \log p + \log \left\{ \frac{\alpha \sqrt{\pi} (\underline{\hat{\omega}}\hat{r} + \hat{\beta} - \hat{\Lambda})}{2\sqrt{\underline{\hat{\omega}}\hat{r}}(1 - \alpha)} \right\} \right), \quad \text{and} \\ \hat{\beta}, \hat{r} \text{ and } \underline{\hat{\omega}} &\text{ are given by (4.5), then,} \end{aligned}$$

$$\text{mFDR} \leq \alpha \quad \text{and} \quad \text{mFNR} \leq L_p p^{-\beta - (\sqrt{\underline{\omega}r} - \sqrt{\beta})^2}.$$

Although two threshold levels  $s$  and  $q$  are not explicitly specified, simulation studies show that the performance of the proposed procedure is insensitive to  $(s, q)$  as long as they are properly chosen from two intervals separated by  $\beta \in (0, 1)$ .

The optimality of the proposed DATE is established for the signal in the partial recovery region with  $\underline{\omega}r < (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$ . If  $\underline{\omega}r \geq (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$ , the region is the full recovery region. The lower bounds of the weighted risk  $H(\Lambda)$  and the mFNR corresponding to this region converge to zero as  $r$  tends to infinity at

each fixed large value of  $p$  as shown in Theorems 1 and 2. However, even when  $\underline{\omega}r \geq (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$ , the upper bounds for these two rates corresponding to the full recovery region will not vanish, since the proposed DATE procedure involves data transformation, precision matrix and tuning parameters estimation each of which contributes non-negligible error at the order of  $o(p^{-1})$ . Although this error is very small, it becomes prominent and dominant as  $r$  is big enough to make two upper bounds established in Theorems 3, 4 and 5 smaller order of  $o(p^{-1})$ , and consequently the upper bounds of the weighted risk  $H(\Lambda)$  and the mFNR will be at the rate of  $o(p^{-1})$ .

## 5. SIMULATION STUDY

Simulation studies were conducted to demonstrate the performance of the proposed procedure for signals recovery under different combinations of signal sparsity controlled by  $\beta$ , signal strength  $r$  and data dependence. The proposed procedure is denoted by DATE $_{\Omega}$  if  $\Omega$  is known and DATE $_{\hat{\Omega}}$  if  $\Omega$  is unknown. For comparison, the BH procedure was also implemented as follows: each of  $p$  coordinates is tested by the two-sample t test to obtain the ordered  $p$ -values  $P_{(1)} < \dots < P_{(p)}$ . Based on the cutoff value  $m = \max\{1 \leq k \leq p : P_{(k)} \leq k\alpha/p\}$ , the coordinates with  $P_i \leq P_{(m)}$  are treated as signal bearing dimensions.

The random samples  $\{X_{ij}\}$  were generated from  $N(\mu_i, \Sigma)$  for  $i = 1, 2$ . Without loss of generality,  $\mu_1 = 0$  and  $\mu_2$  had  $[p^{1-\beta}]$  nonzero coordinates which were uniformly and randomly drawn from  $\{1, \dots, p\}$ . The magnitude of each nonzero entry of  $\mu_2$  was randomly drawn from the interval  $[\sqrt{r \log p/n}, \sqrt{3r \log p/n}]$  and then multiplied by a random sign. Four models were considered for the covariance matrix  $\Sigma = (\sigma_{ij})$ :

(a). AR(1) model:  $\sigma_{ij} = \rho^{|i-j|}$  for  $1 \leq i, j \leq p$ .

(b). Block diagonal model:  $\sigma_{ii} = 1$  for  $i = 1, \dots, p$ , and  $\sigma_{ij} = 0.6$  for  $2(k-1) +$

$1 \leq i \neq j \leq 2k$  where  $k = 1, \dots, \lfloor p/2 \rfloor$ .

(c). Penta-diagonal model:  $\sigma_{ii} = 1$  for  $i = 1, \dots, p$ ,  $\sigma_{ij} = 0.5$  for  $|i - j| = 1$  and  $\sigma_{ij} = 0.2$  for  $|i - j| = 2$ .

(d). Random sparse matrix model: first generate a  $p \times p$  matrix  $\Gamma$  each row of which has only one non-zero element that is randomly chosen from  $\{1, \dots, p\}$  with magnitude generated from  $\text{Unif}(1, 2)$  multiplied by a random sign.  $\Sigma$  is then obtained by standardizing  $\Gamma\Gamma^T + \mathbf{I}$  to have unit diagonal elements.

To apply the  $\text{DATE}_{\hat{\Omega}}$ , we need to estimate  $\Omega$ . For models (a) – (c), the Cholesky decomposition approach (Bickel and Levina, 2008a) was implemented. Recall that the precision matrix  $\Omega$  can be decomposed as  $\Omega = (I - A)'D^{-1}(I - A)$  where  $A$  is a lower triangular matrix with zero diagonals and  $D$  is a diagonal matrix. The elements below the diagonal element on the  $k$ th row of  $A$  can be thought as the regression coefficients of the  $k$ th component on its predecessors, and the  $k$ th diagonal element of  $D$  is the corresponding residual variance. Let  $A_\tau$  be the  $\tau$ -banded lower triangular matrix of  $A$  and  $D_\tau$  be the corresponding residual variances on the diagonals. The  $\tau$ -banded precision matrix  $\Omega_\tau = (I - A_\tau)'D_\tau^{-1}(I - A_\tau)$ . Given a sample,  $A_\tau$  and  $D_\tau$  can be estimated by the least square estimation, which leads to

$$\hat{\Omega}_\tau = (I - \hat{A}_\tau)' \hat{D}_\tau^{-1} (I - \hat{A}_\tau),$$

where the banding width parameter  $\tau$  in the estimation of  $\Omega$  was chosen according to the data-driven procedure proposed by Bickel and Levina (2008a). For a given data set, we divided it into two subsamples by repeated ( $N = 50$  times) random data split. For the  $l$ -th split,  $l \in \{1, \dots, N\}$ , we let  $\hat{\Sigma}_\tau^{(l)} = \{(I - \hat{A}_\tau^{(l)})'\}^{-1} \hat{D}_\tau^{(l)} (I - \hat{A}_\tau^{(l)})^{-1}$  be the Cholesky decomposition of  $\Sigma$  obtained from the first subsample by taking the same approach described in previous section for  $\hat{A}_\tau^{(l)}$  and  $\hat{D}_\tau^{(l)}$ . Also we let  $S_n^{(l)}$  be the

sample covariance matrix obtained from the second subsample. Then the banding parameter  $\tau$  is selected as

$$\hat{\tau} = \min_{\tau} \frac{1}{N} \sum_{l=1}^N \|\hat{\Sigma}_{\tau}^{(l)} - S_n^{(l)}\|_F, \quad (5.1)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm.

In model (d),  $\Sigma$  is first estimated by applying the thresholding operator  $T_m$  to the sample covariance matrix  $S_n$  such that

$$T_m(S_n) = [s_{ij} \mathbf{I}(|s_{ij}| \geq m)].$$

As shown by Bickel and Levina (2008b),  $\|T_m(S_n) - \Sigma\| = o_p(1)$  under the condition (C4) where  $\|\cdot\|$  is the spectral norm. The threshold  $m$  can be selected by the cross-validation method. Given a data of size  $n$ , we split it into two sub-samples with sizes of  $n_1 = n\{1 - \frac{1}{\log(n)}\}$  and  $n_2 = \frac{n}{\log(n)}$  for  $N = 50$  times. For the  $l$ -th split, let  $S_{1,l}$  and  $S_{2,l}$  be the sample covariance matrices based on the  $n_1$  and  $n_2$  observations. The threshold  $\hat{m}$  is chosen to minimize

$$R_s = \frac{1}{N} \sum_{l=1}^N \|T_m(S_{1,l}) - S_{2,l}\|_F^2. \quad (5.2)$$

Then  $\Omega$  can be estimated by  $T_{\hat{m}}^{-1}(S_n)$  since  $T_{\hat{m}}(S_n)$  is positive definite provided that  $\|T_{\hat{m}} - T_0\| \leq \epsilon$  and  $\lambda_{\min} > \epsilon$  (Bickel and Levina, 2008b).

The performance of each signal recovery procedure was evaluated by mFDR, mFNR and the average number of true positives ATP based on 100 replications. The nominal FDR level was set at  $\alpha = 0.05$ . Figure 3 displays the performance of three procedures with different values of signal strength  $r$  and data dependence  $\rho$  under model (a) when  $p = 500$ . In the first row of the Figure, data were weakly dependent and all three procedures had the mFDR controlled below the nominal level 0.05 except  $r = 0.4$ . The distortion of the mFDR at  $r = 0.4$  is due to the fact that the signals fall into the region of no recovery since  $r < \beta/\bar{\omega}$  with  $\bar{\omega} = 1.08$  when  $\rho = 0.2$ .

With the dependence increased from  $\rho = 0.2$  to  $0.6$ , the inflation in mFDR was mitigated since  $r > \beta/\underline{\omega}$  with  $\underline{\omega} = 1.56$  when  $\rho = 0.6$ . Although the  $\text{DATE}_\Omega$ ,  $\text{DATE}_{\hat{\Omega}}$  and BH performed similarly in terms of the mFNR and ATP with weakly dependent  $\rho = 0.2$ , both  $\text{DATE}_\Omega$  and  $\text{DATE}_{\hat{\Omega}}$  had more ATP which is close to the number of true signals  $[500^{0.4}] = 12$  for strong signal strength  $r$ , and suffered less mFNR than the BH with moderate dependent  $\rho = 0.6$ , which confirms that the data dependence is utilized by the proposed procedures for signal identification. When dimension  $p$  was increased from 500 to 1000, Figure 4 demonstrates the results similar to Figure 3. Specially with strong signal strength  $r$ , the recovery of signals by both  $\text{DATE}_\Omega$  and  $\text{DATE}_{\hat{\Omega}}$  was close to the number of true signals  $[1000^{0.4}] = 16$ .

The performance of three procedures with various dependent structures defined in models (b)-(d) were also displayed in Figures 5-7. Again, both  $\text{DATE}_\Omega$  and  $\text{DATE}_{\hat{\Omega}}$  performed better than the BH in terms of mFNR and ATP even though all the procedures had the mFDR controlled at the nominal level 0.05.

$\text{DATE}_\Omega$  depends on the level of threshold  $s$  and  $\text{DATE}_{\hat{\Omega}}$  depends on both  $s$  and  $q$ , which are required to be chosen from intervals  $(0, \beta)$  and  $(\beta, \underline{\omega}r)$  respectively. Table 1 displays the performance of both  $\text{DATE}_\Omega$  and  $\text{DATE}_{\hat{\Omega}}$  in terms of mFDR and mFNR with different values of  $s$  and  $q$  under model (a) where  $\beta = 0.6$ ,  $\rho = 0.6$  and  $r = 0.8$ . As we can see, the proposed procedure is insensitive to the choice of  $s$  and  $q$  as long as they are chosen properly from the intervals.

## 6. EMPIRICAL STUDY

We applied the proposed DATE procedure to a human breast cancer dataset which is available at <http://www.ncbi.nlm.nih.gov>. The data were analyzed by Richardson et al. (2006) to provide insight into the molecular pathogenesis of Sporadic basal-like cancers (BLC) that is a distinct class of human breast cancers. As discussed by

Richardson et al. (2006), BLC specimens display X chromosome abnormalities in the sense that most of the BLC cases lack markers of a normal inactive X chromosome, which are rare in non-BLC specimens. So our interest on this data set is to display these X chromosome abnormalities by identifying the differentially expressed genes between the BLC and non-BLC. For this purpose, we formed two samples by taking 18 sporadic BLC specimens and 20 non-BLC specimens from the original data, and each sample contains 1438 genes obtained from chromosome X.

To apply the DATE procedure, we first estimated  $\Omega$  in (1.2) where  $\Sigma_1 \neq \Sigma_2$  in general. To facilitate a simpler estimation, we changed the two-sample problem into an one-sample problem by defining

$$Y_i = X_{1i} - \sqrt{\frac{n_1}{n_2}} X_{2i} + \frac{1}{\sqrt{n_1 n_2}} \sum_{j=1}^{n_1} X_{2j} - \frac{1}{n_2} \sum_{l=1}^{n_2} X_{2l} \quad i = 1, \dots, n_1,$$

where we assume  $n_1 \leq n_2$ . It can be shown that  $Y_i \stackrel{i.i.d.}{\sim} N(\delta, \Sigma_w)$  where  $\Sigma_w = \Sigma_1 + \frac{n_1}{n_2} \Sigma_2$  under the model (1.1). Note that  $\Omega = \frac{n_1 + n_2}{n_2} \Sigma_w^{-1}$ . To estimate  $\Omega$ , we only need to estimate  $\Sigma_w^{-1}$  based on  $Y_i$  for  $i = 1, \dots, n_1$ . The available packages for this purpose include *glasso*, *Covpath* and *CLIME*, which are coded based on different estimation approaches discussed in Section 3. To implement a fast algorithm, we adopted the *glasso* which chooses the non-negative definite matrix  $\hat{\Omega}_{Glasso}$  to maximize a  $L_1$ -regularized log-likelihood:

$$\log \det(\Sigma^{-1}) - \text{tr}(S \Sigma^{-1}) - \rho \|\Sigma^{-1}\|_1,$$

where  $S$  is the sample covariance matrix and  $\rho$  is a tuning parameter controlling the  $L_1$  shrinkage. To select the regularization parameter  $\rho$ , we considered the package *huge* developed by Zhao, Liu, Roeder, Lafferty and Wasserman (2012) where three methods are provided: the stability approach for regularization selection, rotation information criterion and a likelihood-based extended Bayesian information criterion. Except the DATE procedure, we also considered the classical BH procedure integrated

with two-sample t test as a comparison.

In order to identify the differentially expressed genes, the FDR was chosen to be controlled at  $\alpha = 0.001, 0.005$  and  $0.01$ . Table 2 summarizes the number of differentially expressed genes identified by the BH only and the DATE only, and both procedures. By carefully investigating the genes identified by both procedures, we found that the XIST (X inactive specific transcript) gene was discovered. This gene is in charge of an early developmental process in females and provides dosage equivalence between males and females. The XIST difference is thought as one of the characteristics for the BLC according to Richardson et al. (2006). Moreover, the authors argue that there exists the overexpression of a small subset of genes on chromosome X for BLC. In Table 3, we list additional 17 genes that are identified by the DATE but missed by the BH with the FDR controlled at  $\alpha = 0.001$ . The association of these genes with the BLC may deserve some further biological investigation.

## 7. DISCUSSION

Signal identification is different from its closely related problem of signal detection. Whereas the detection focuses purely on the presence of signals, the signal identification is designated for locating the signals. The advantage of dependence for signal detection was exploited by Hall and Jin (2010) who showed that the detection boundary can be lowered by incorporating the data correlation. However, it is unclear that the similar advantage can be offered by data dependence for signal identification. The current paper attempts to answer this question. Our analysis shows that both full and partial signal identification boundaries for dependent data are lower than those without dependence. Our result, combined with the findings in Hall and Jin (2010), shows that data dependence is advantageous in both signal detection and signal identification.

When data dependence is present, it becomes challenging to find a procedure which minimizes the FNR while controlling the FDR at a pre-specified level  $\alpha < 1$ . When both signals and precision matrix are sparse, the proposed DATE procedure takes advantage of dependence through the transformation to enhance the signal strength and is shown to have the faster convergence rate in mFNR than other procedures without take data dependence into account. The current work is related with that of Ji and Jin (2012) and of Ji and Zhao (2014), where the authors considered the variable selection and multiple testing in the high dimensional regression problem. In our paper, the precision matrix for data transformation is assumed to be sparse. More research is needed to develop an optimal procedure for signal identification under general dependence structure.



## APPENDIX: TECHNICAL DETAILS.

### A.1. Lemmas 1 and 2

**Lemma 1.** For any positive definite matrix  $A_{p,p} = (a_{ij})_{p \times p}$  and its inverse  $B_{p,p} = (b_{ij})_{p \times p}$ , the following inequality holds

$$a_{ii} \cdot b_{ii} \geq 1 \quad i = 1, \dots, p.$$

Proof. We first show that  $a_{pp} \cdot b_{pp} \geq 1$ . To this end, we write

$$A_{p,p} = \begin{pmatrix} A_{p-1,p-1} & a_{p-1,1} \\ a'_{p-1,1} & a_{pp} \end{pmatrix}.$$

Then using the result from matrix inversion in block form, we have

$$b_{pp} = (a_{pp} - a'_{p-1,1} A'_{p-1,p-1} a_{p-1,1})^{-1}, \quad (\text{A.1})$$

which implies that  $a_{pp} \cdot b_{pp} \geq 1$  since  $a'_{p-1,1} A'_{p-1,p-1} a_{p-1,1} \geq 0$ .

For any  $i$ , we can switch  $a_{ii}$  from its original position to the position  $(p, p)$  using the permutation matrix  $P_{p,p}$ . Accordingly,  $b_{ii}$  is moved from its original location to  $(p, p)$  by the same matrix  $P_{p,p}$ . By the fact that the permutation matrix is also the orthogonal matrix, we have

$$P_{p,p} A_{p,p} P_{p,p} P_{p,p} B_{p,p} P_{p,p} = I_{p,p}.$$

Therefore, from (A.1), we have  $a_{ii} \cdot b_{ii} \geq 1$  for any  $i$ . This completes the proof of Lemma 1.

For any  $k \in \{1, \dots, p\}$ , we let

$$A_k(\Omega) = \{l : 1 \leq l \leq p, |\omega_{kl}| \geq L_p^{-1}\},$$

and  $B_k$  be the event that  $\{\delta_l = 0 \text{ for all } l \neq k \text{ and } l \in A_k\}$ . If  $\Omega$  is unknown, it can be estimated by  $\hat{\Omega}$  (Cai, Liu and Luo (2011)), which, with probability  $1 - O(p^{-\tau})$  where

$\tau$  is a positive constant, satisfies

$$\|\hat{\Omega} - \Omega\|_{L_1} = O_p\left\{\left(\frac{\log p}{n}\right)^{\frac{1-\zeta}{2}}\right\}.$$

Then, let  $D_p$  be the event  $\{\|\hat{\Omega} - \Omega\|_{L_1} \leq (\frac{\log p}{n})^{\frac{1-\zeta}{2}}\}$ .

**Lemma 2.** Assume conditions (C2)-(C4). Over the event  $\{\delta_k \neq 0\} \cap B_k \cap D_k$ ,

$$\hat{\delta}_{\Omega,k} = \omega_{kk}\delta_k + o(n^{-1/2}).$$

Proof: We first consider that  $\Omega$  is known. By condition (C3), the number of elements in set  $A_k(\Omega)$  satisfies that  $|A_k(\Omega)| \leq ML_p$ . Since  $\beta > 1/2$ , condition (C2) leads to

$$P(\delta_k \neq 0, B_k^c) \leq \sum_{l \in A_k, l \neq k} P(\delta_k \neq 0, \delta_l \neq 0) \leq ML_p p^{-2\beta} = o(p^{-1}). \quad (\text{A.2})$$

Note that  $\delta_{\Omega,k} = \sum_{l \in A_k} \omega_{kl}\delta_l + \sum_{l \in A_k^c} \omega_{kl}\delta_l$ . Over the event  $\{\delta_k \neq 0\} \cap B_k$ ,  $\sum_{l \in A_k} \omega_{kl}\delta_l = \omega_{kk}\delta_k$ . Moreover, for  $l \in A_k^c$ ,  $|\omega_{kl}|^{\zeta-1}/L_p^{1-\zeta} > 1$ . Therefore, using condition (C3) again, for some constant  $c$ , we have

$$|\delta_{\Omega,k} - \omega_{kk}\delta_k| \leq \max_l |\delta_l| L_p^{\zeta-1} \sum_{l \in A_k^c} |\omega_{kl}|^{\zeta} \leq \max_l |\delta_l| L_p^{\zeta-1} c.$$

Since  $|\delta_l| \sim \sqrt{2r \log p / n}$  and  $\zeta < 1$ , we can choose a large enough slowly varying function  $L_p$  such that  $\max_l |\delta_l| L_p^{\zeta-1} = o(n^{-1/2})$ . Thus, we have  $\delta_{\Omega,k} = \omega_{kk}\delta_k + o(n^{-1/2})$ .

Next, we consider that  $\Omega$  is unknown. If  $\tau > 1$ ,  $P(D_p^c) = o(p^{-1})$  by the definition of the event of  $D_p$ . Note that  $\hat{\delta}_{\Omega,k} = \delta_{\Omega,k} + \{(\hat{\Omega} - \Omega)\delta\}_k$ . Then over the event  $D_p$  and by condition (C4),

$$\{(\hat{\Omega} - \Omega)\delta\}_k \leq \max_l |\delta_l| \cdot \|\hat{\Omega} - \Omega\|_{L_1} \leq \left(\frac{\log p}{n}\right)^{1-\frac{\zeta}{2}} = o(n^{-1/2}).$$

Then, over the event  $\{\delta_k \neq 0\} \cap B_k \cap D_k$ ,  $\hat{\delta}_{\Omega,k} = \omega_{kk}\delta_k + o(n^{-1/2})$ . This completes the proof of Lemma 2.

## A.2. Proof of Lemma 3

Recall that in Lemma 2,  $B_k$  is the event that  $\{\delta_l = 0 \text{ for all } l \neq k \text{ and } l \in A_k\}$ . Since  $\underline{\omega}r < (\sqrt{1-\Lambda} + \sqrt{1-\beta})^2$ , it can be shown that  $\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r) < 1$ . Also using (A.2), we know that it is sufficient to prove Lemma 3 over the event  $B_k$ . Without loss of generality, we assume  $\sqrt{n}\delta_k = \sqrt{2r\log p}$ . The result for negative signals can be derived similarly. Note that  $s < (\underline{\omega}r + \beta - \Lambda)^2/(4\underline{\omega}r) = \{(\underline{\omega}r + \beta - \Lambda)^2/(2\underline{\omega}r)^2\}(\underline{\omega}r) < \underline{\omega}r$  since  $\underline{\omega}r > \beta - \Lambda$ . Then, over the event  $B_k$ ,

$$\begin{aligned}
& \mathbb{P}(T_k < 2s\log p, \delta_k \neq 0) \\
& \leq \mathbb{P}(\delta_k \neq 0)\mathbb{P}(T_k < 2s\log p | \delta_k \neq 0 \cap B_k) \\
& = p^{-\beta} \left\{ \mathbb{P}\left(\frac{\sqrt{n}(\bar{Z}_1^{(k)} - \bar{Z}_2^{(k)})}{\sqrt{\omega_{kk}}} - \frac{\sqrt{n}\delta_{\Omega_k}}{\sqrt{\omega_{kk}}} < \sqrt{2s\log p} - \frac{\sqrt{n}\delta_{\Omega_k}}{\sqrt{\omega_{kk}}} | \delta_k \neq 0 \cap B_k\right) \right. \\
& \quad \left. - \mathbb{P}\left(\frac{\sqrt{n}(\bar{Z}_1^{(k)} - \bar{Z}_2^{(k)})}{\sqrt{\omega_{kk}}} - \frac{\sqrt{n}\delta_{\Omega_k}}{\sqrt{\omega_{kk}}} < -\sqrt{2s\log p} - \frac{\sqrt{n}\delta_{\Omega_k}}{\sqrt{\omega_{kk}}} | \delta_k \neq 0 \cap B_k\right) \right\} \\
& \leq p^{-\beta} L_p p^{-(\sqrt{\omega_{kk}r} - \sqrt{s})^2} \{1 + o(1)\} \\
& \leq p^{-\beta} L_p p^{-(\sqrt{\underline{\omega}r} - \sqrt{s})^2} \{1 + o(1)\}.
\end{aligned}$$

Since  $s < (\underline{\omega}r + \beta - \Lambda)^2/(4\underline{\omega}r)$ , we have

$$\sum_{k=1}^p \mathbb{P}(T_k < 2s\log p, \delta_k \neq 0) \leq L_p p^{1 - \{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)}.$$

Next we consider that  $\Omega$  is unknown. Let  $D_p$  be the event

$$\left\{ \max_{1 \leq k \leq p} \left| \sum_l (\hat{\Omega}_{kl} - \Omega_{kl})(\bar{X}_1^{(l)} - \bar{X}_2^{(l)}) \right| \leq \left(\frac{\log p}{n}\right)^{1-\frac{\alpha}{2}}, \max_{1 \leq k \leq p} |\hat{\omega}_{kk} - \omega_{kk}| \leq \left(\frac{\log p}{n}\right)^{\frac{1-\alpha}{2}} \right\}.$$

Note that

$$|\hat{T}_k^{\frac{1}{2}}| = \left| \left\{ \frac{\sqrt{n} \sum_l \Omega_{kl}(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})}{\sqrt{\omega_{kk}}} + \frac{\sqrt{n} \sum_l (\hat{\Omega}_{kl} - \Omega_{kl})(\bar{X}_1^{(l)} - \bar{X}_2^{(l)})}{\sqrt{\omega_{kk}}} \right\} \frac{1}{1 + \frac{\sqrt{\omega_{kk}} - \sqrt{\omega_{kk}}}{\sqrt{\omega_{kk}}}} \right|,$$

which by condition (C4), leads to

$$\begin{aligned}
\mathbb{P}(\hat{T}_k < 2s\log p, \delta_k \neq 0) & \leq p^{-\beta} \mathbb{P}(\hat{T}_k < 2s\log p | \delta_k \neq 0 \cap B_k \cap D_p) + \mathbb{P}(\delta_k \neq 0 \cap B_k^c) + \mathbb{P}(D_p^c) \\
& \leq p^{-\beta} \mathbb{P}(T_k < 2s\log p | \delta_k \neq 0) + \mathbb{P}(\delta_k \neq 0 \cap B_k^c) + \mathbb{P}(D_p^c) \\
& \leq L_p p^{\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)} + \mathbb{P}(\delta_k \neq 0 \cap B_k^c) + \mathbb{P}(D_p^c).
\end{aligned}$$

Since  $P(\delta_k \neq 0 \cap B_k^c) = o(p^{-1})$  by (A.2),  $P(D_p^c) = o(p^{-1})$  and  $\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r) < 1$ ,

$$P(\hat{T}_k < 2s \log p, \delta_k \neq 0) \leq L_p p^{-\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)} \{1 + o(1)\}.$$

This completes the proof of Lemma 3.

### A.3. Proof of Lemma 4

Similar to Lemma 2.2 of Ji and Jin (2012), it can be shown that with probability  $1 - o(p^{-1})$ , each row of the regularized  $\Omega^*$  defined by (4.3) has no more than  $L_p$  nonzero components and also  $\|\Omega - \Omega^*\|_{L_1} \leq C(\log p)^{-(1-\alpha)}$ .

Note that  $\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r) < 1$ . And a connected graph with size  $l + 1$  for  $l \geq 1$  always contains a connected graph with size  $l$ . Then with respect to  $(V_0, \Omega^*)$ , it is sufficient to show that there exists an integer  $m$  such that

$$P\{\mathcal{U}(s) \text{ contains a connected graph with size } m\} \leq o(p^{-1}).$$

Recall that  $D_p$  is the event

$$\left\{ \max_{1 \leq k \leq p} \left| \sum_l (\hat{\Omega}_{kl} - \Omega_{kl})(\bar{X}_1^{(l)} - \bar{X}_2^{(l)}) \right| \leq \left(\frac{\log p}{n}\right)^{1-\frac{\alpha}{2}}, \max_{1 \leq k \leq p} |\hat{\omega}_{kk} - \omega_{kk}| \leq \left(\frac{\log p}{n}\right)^{\frac{1-\alpha}{2}} \right\}.$$

And  $P(D_p^c) = o(p^{-1})$ . Therefore, it is enough to show that

$$P\{\mathcal{U}(s) \text{ contains a connected graph with size } m, D_p\} \leq o(p^{-1}).$$

Since there are logarithmically large number of nonzero elements in each row or column of  $\Omega^*$ , there are at most  $pL_p^m$  connected graphs with size  $m$  by the result from Frieze and Molloy (1999). As a result, by the union bound, it is sufficient to show that for any connected graph of size  $m$ , say  $\mathcal{I} = \{k_1, \dots, k_m\}$ ,

$$P\{\mathcal{I} \subset \mathcal{U}(s), D_p\} \leq o(p^{-2}).$$

Let  $\hat{T} = \{\hat{T}_1, \dots, \hat{T}_p\}$  and  $\mathbf{1}_p = \{1, \dots, 1\}$ . Then we need to show that

$$\mathbb{P}\{\hat{T}^{\mathcal{I}} \geq 2s\log p \mathbf{1}_p^{\mathcal{I}}, D_p\} \leq o(p^{-2}).$$

Since for any  $1 \leq k \leq p$ ,

$$\hat{T}_k = \left\{ \frac{\sqrt{n} \sum_l \Omega_{kl} (\bar{X}_1^{(l)} - \bar{X}_2^{(l)})}{\omega_{kk}^{1/2}} + \frac{\sqrt{n} \sum_l (\hat{\Omega}_{kl} - \Omega_{kl}) (\bar{X}_1^{(l)} - \bar{X}_2^{(l)})}{\omega_{kk}^{1/2}} \right\}^2 \left( \frac{1}{1 + \frac{\hat{\omega}_{kk} - \omega_{kk}}{\omega_{kk}}} \right),$$

we only need to show that

$$\mathbb{P}\{|(T^{\mathcal{I}})^{1/2}| \geq (2s\log p)^{1/2} \mathbf{1}_p^{\mathcal{I}}, D_p\} \leq o(p^{-2}).$$

Let  $\tilde{\delta}_{\Omega}^{\mathcal{I}} = (\delta_{\Omega, k_1}/\omega_{k_1 k_1}^{1/2}, \dots, \delta_{\Omega, k_m}/\omega_{k_m k_m}^{1/2})$  and  $\mathcal{E} = (T^{\mathcal{I}})^{1/2} - \sqrt{n} \tilde{\delta}_{\Omega}^{\mathcal{I}}$ . Then it can be shown that

$$\mathcal{E} \sim \mathcal{N}(0, \bar{\Omega}^{\mathcal{I}, \mathcal{I}}),$$

where  $\bar{\Omega}_{ij} = \omega_{ij}/(\omega_{ii}\omega_{jj})^{1/2}$  for  $i, j \in \mathcal{I}$ . By Cauchy-Schwartz inequality,

$$\|\mathcal{E}\|^2 \geq \frac{1}{2} \| (T^{\mathcal{I}})^{1/2} \|^2 - \|\sqrt{n} \tilde{\delta}_{\Omega}^{\mathcal{I}}\|^2. \quad (\text{A.3})$$

Since the largest eigenvalue of  $\bar{\Omega}^{\mathcal{I}, \mathcal{I}}$  is not greater than that of  $\bar{\Omega}$ . The latter has the largest eigenvalue no greater than  $C_0/\underline{\omega} \leq C_0$  since  $\underline{\omega} \geq 1$ . Therefore,

$$\mathcal{E}'(\bar{\Omega}^{\mathcal{I}, \mathcal{I}})^{-1} \mathcal{E} \geq \frac{1}{C_0} \|\mathcal{E}\|^2. \quad (\text{A.4})$$

Moreover, by the construction of the thresholding,

$$\| (T^{\mathcal{I}})^{1/2} \|^2 \geq 2ms\log p. \quad (\text{A.5})$$

Combining (A.3), (A.4) and (A.5), we have

$$\mathcal{E}'(\bar{\Omega}^{\mathcal{I}, \mathcal{I}})^{-1} \mathcal{E} \geq \frac{1}{C_0} \{ms\log p - \|\sqrt{n} \tilde{\delta}_{\Omega}^{\mathcal{I}}\|^2\}.$$

Using Lemma A.3 of Ji and Jin (2012), we can show that

$$\mathbb{P}(\|\sqrt{n} \tilde{\delta}_{\Omega}^{\mathcal{I}}\|^2 \geq (C_0 k + cm(\log p)^{-2(1-\alpha)})(2s\log p), D_p) \leq L_p p^{-\beta k},$$

where  $k$  is chosen to satisfy  $(C_0k + cm(\log p)^{-2(1-\alpha)})(2s\log p) \leq \frac{1}{2}ms\log p$ . Denote  $A_p$  to be the event

$$\{||\sqrt{n}\tilde{\delta}_\Omega^\mathcal{I}||^2 \geq (C_0k + cm(\log p)^{-2(1-\alpha)})(2s\log p)\}.$$

Then, we have  $P(A_p \cap D_p) \leq L_p p^{-\beta k}$ . Then, for sufficiently large  $m$ ,

$$\begin{aligned} P\{\mathcal{I} \subset \mathcal{U}(s), D_p\} &\leq P\{\mathcal{E}'(\bar{\Omega}^{\mathcal{I}\mathcal{I}})^{-1}\mathcal{E} \geq \frac{1}{2C_0}(ms\log p)\} + P(A_p \cap D_p) \\ &\leq L_p(p^{-\frac{1}{4C_0}ms} + p^{-\beta k}) \\ &= o(p^{-2}). \end{aligned}$$

This completes the proof of Lemma 4.

#### A.4. Proof of Theorem 1

To make the discussion earlier, we change the two-sample problem into an one-sample problem. Without loss of generality, we assume  $n_1 \leq n_2$  and define

$$Y_i = X_{1i} - \sqrt{\frac{n_1}{n_2}}X_{2i} + \frac{1}{\sqrt{n_1n_2}} \sum_{j=1}^{n_1} X_{2j} - \frac{1}{n_2} \sum_{l=1}^{n_2} X_{2l} \quad i = 1, \dots, n_1. \quad (\text{A.6})$$

It can be shown that  $Y_i \stackrel{i.i.d.}{\sim} N(\delta, \Sigma_1 + \frac{n_1}{n_2}\Sigma_2)$  for  $i = 1, \dots, n_1$  under the model (1.1).

Note that the loss function for  $j$ th dimension is  $L(\theta_j, \hat{\theta}_j) = \theta_j(1 - \hat{\theta}_j) + p^{-\Lambda}(1 - \theta_j)\hat{\theta}_j$  where  $\theta_j = 0$  if  $\delta_j = 0$  and  $\theta_j = 1$  otherwise, and  $\hat{\theta}_j$  is the decision rule with value equal to either 0 or 1. Clearly,  $L(\theta, a) = \sum_j L(\theta_j, \hat{\theta}_j)$ . The following derivation for  $j$ th dimension can be also extended to other dimensions. Therefore, without making any confusion, we drop the subscript  $j$ . Let  $\tilde{\delta} = \delta - \alpha e_j$  where  $e_j$  is a  $p \times 1$  vector with  $j$ th element equal to 1. Let  $h(\mathcal{Y}; \tilde{\delta}, \alpha)$  be the joint density of  $(Y_1, \dots, Y_{n_1})$  where

$Y_i$  is defined in (A.6):

$$\begin{aligned}
h(\mathcal{Y}; \tilde{\delta}, \alpha) &= (2\pi)^{-n_1 p/2} |\tilde{\Sigma}|^{-n_1/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n_1} \{(Y_i - \tilde{\delta})' \tilde{\Sigma}^{-1} (Y_i - \tilde{\delta})\}\right) \\
&\quad \exp\left\{\alpha e'_j \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta}) - n_1/2\alpha^2 \tilde{\omega}_{jj}\right\} \\
&= h(\mathcal{Y}; \tilde{\delta}, 0) \exp\left\{\alpha e'_j \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta}) - n_1/2\alpha^2 \tilde{\omega}_{jj}\right\},
\end{aligned}$$

where  $\tilde{\Sigma} = (\Sigma_1 + n_1/n_2 \Sigma_2)$  and  $\tilde{\omega}_{jj}$  is the  $j$ th diagonal element of  $\tilde{\Sigma}^{-1}$ . If we let

$$f_0(\mathcal{Y}) = \int h(\mathcal{Y}; \tilde{\delta}, 0) dF(\tilde{\delta}) \quad \text{and} \quad f_1(\mathcal{Y}) = \int h(\mathcal{Y}; \tilde{\delta}, \alpha) d\pi_p(\alpha) dF(\tilde{\delta}), \quad (\text{A.7})$$

where  $F(\tilde{\delta})$  is the joint CDF of  $\tilde{\delta}$  and  $\pi_p(\alpha)$  is the CDF of  $\alpha$  defined in condition (C2). Then the following Bayesian decision rule minimizes the risk function for  $j$ th dimension:

$$\hat{\theta}_j = \mathbf{I}\left\{\frac{(1 - \epsilon_p)f_0(\mathcal{Y})}{\epsilon_p f_1(\mathcal{Y})} \leq p^\Lambda\right\},$$

where  $\epsilon_p = p^{-\beta}$ . The corresponding risk function is

$$\begin{aligned}
H_j &= \mathbf{E}\{L(\theta_j, \hat{\theta}_j)\} \\
&= \mathbf{E}(\mathbf{E}\{L(\theta_j, \hat{\theta}_j)\}|\mathcal{Y}) \\
&= \mathbf{E}\left\{\frac{\epsilon_p f_1(\mathcal{Y})}{(1 - \epsilon_p)f_0(\mathcal{Y}) + \epsilon_p f_1(\mathcal{Y})}(1 - \hat{\theta}_j) + p^{-\Lambda} \frac{(1 - \epsilon_p)f_0(\mathcal{Y})}{(1 - \epsilon_p)f_0(\mathcal{Y}) + \epsilon_p f_1(\mathcal{Y})} \hat{\theta}_j\right\} \\
&= \int_{A^c} \epsilon_p f_1(\mathcal{Y}) d\mathcal{Y} + p^{-\Lambda} \int_A (1 - \epsilon_p)f_0(\mathcal{Y}) d\mathcal{Y} \\
&= \epsilon_p - \int_A |p^{-\Lambda}(1 - \epsilon_p)f_0(\mathcal{Y}) - \epsilon_p f_1(\mathcal{Y})| d\mathcal{Y},
\end{aligned}$$

where the set  $A = \{\mathcal{Y} : \hat{\theta}_j = 1\}$ , and from line two to line three, we have used the fact that

$$\mathbf{E}(\theta_j|\mathcal{Y}) = \mathbf{P}(\theta_j = 1|\mathcal{Y}) = \frac{\mathbf{P}(\mathcal{Y}|\theta_j = 1)\mathbf{P}(\theta_j = 1)}{\mathbf{P}(\mathcal{Y}|\theta_j = 0)\mathbf{P}(\theta_j = 0) + \mathbf{P}(\mathcal{Y}|\theta_j = 1)\mathbf{P}(\theta_j = 1)}.$$

Similarly,

$$\begin{aligned}
H_j &= \int_{A^c} \epsilon_p f_1(\mathcal{Y}) d\mathcal{Y} + p^{-\Lambda} \int_A (1 - \epsilon_p)f_0(\mathcal{Y}) d\mathcal{Y} \\
&= p^{-\Lambda}(1 - \epsilon_p) - \int_{A^c} |p^{-\Lambda}(1 - \epsilon_p)f_0(\mathcal{Y}) - \epsilon_p f_1(\mathcal{Y})| d\mathcal{Y}.
\end{aligned}$$

Then the following result can be derived:

$$H_j = \frac{1}{2} \left\{ p^{-\Lambda}(1 - \epsilon_p) + \epsilon_p - \int_A |p^{-\Lambda}(1 - \epsilon_p)f_0 - \epsilon_p f_1| d\mathcal{Y} \right\}, \quad (\text{A.8})$$

where, by Fubini's Theorem,

$$\begin{aligned} & \int_A |p^{-\Lambda}(1 - \epsilon_p)f_0 - \epsilon_p f_1| d\mathcal{Y} \\ &= \int \left| \int \{p^{-\Lambda}(1 - \epsilon_p)h(\mathcal{Y}; \tilde{\delta}, 0) - \epsilon_p h(\mathcal{Y}; \tilde{\delta}, \alpha)\} d\pi_p(\alpha) dF(\tilde{\delta}) \right| d\mathcal{Y} \\ &\leq \int H(\tilde{\delta}, \alpha) d\pi_p(\alpha) dF(\tilde{\delta}), \end{aligned}$$

where  $H(\tilde{\delta}, \alpha) = \int |p^{-\Lambda}(1 - \epsilon_p)h(\mathcal{Y}; \tilde{\delta}, 0) - \epsilon_p h(\mathcal{Y}; \tilde{\delta}, \alpha)| d\mathcal{Y}$ . It can be shown that  $H(\tilde{\delta}, \alpha) = H(\tilde{\delta}, -\alpha)$  and  $H(\tilde{\delta}, \alpha)$  is an increasing function of  $\alpha > 0$ . Hence, for  $\alpha \in [-\tau_p, 0) \cup (0, \tau_p]$  where  $\tau_p = \sqrt{2r \log p / n}$ , we have  $H(\tilde{\delta}, \alpha) \leq H(\tilde{\delta}, \tau_p)$ . As a result,

$$\int_A |p^{-\Lambda}(1 - \epsilon_p)f_0 - \epsilon_p f_1| d\mathcal{Y} \leq \int H(\tilde{\delta}, \tau_p) dF(\tilde{\delta}), \quad (\text{A.9})$$

where, if we let  $D_p = \{\mathcal{Y} : \epsilon_p \exp\{\tau_p e'_j \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta}) - n_1/2\tau_p^2 \tilde{\omega}_{jj}\} > p^{-\Lambda}(1 - \epsilon_p)\}$ , then

$$\begin{aligned} H(\tilde{\delta}, \tau_p) &= - \int_{D_p} \{p^{-\Lambda}(1 - \epsilon_p)h(\mathcal{Y}; \tilde{\delta}, 0) - \epsilon_p h(\mathcal{Y}; \tilde{\delta}, \tau_p)\} d\mathcal{Y} \\ &\quad + \int_{D_p^c} \{p^{-\Lambda}(1 - \epsilon_p)h(\mathcal{Y}; \tilde{\delta}, 0) - \epsilon_p h(\mathcal{Y}; \tilde{\delta}, \tau_p)\} d\mathcal{Y}. \end{aligned}$$

This, together with the fact that

$$\begin{aligned} p^{-\Lambda}(1 - \epsilon_p) + \epsilon_p &= \int_{D_p} \{p^{-\Lambda}(1 - \epsilon_p)h(\mathcal{Y}; \tilde{\delta}, 0) + \epsilon_p h(\mathcal{Y}; \tilde{\delta}, \tau_p)\} d\mathcal{Y} \\ &\quad + \int_{D_p^c} \{p^{-\Lambda}(1 - \epsilon_p)h(\mathcal{Y}; \tilde{\delta}, 0) + \epsilon_p h(\mathcal{Y}; \tilde{\delta}, \tau_p)\} d\mathcal{Y}, \end{aligned}$$

leads to

$$H(\tilde{\delta}, \tau_p) = p^{-\Lambda}(1 - \epsilon_p) + \epsilon_p - 2 \left\{ p^{-\Lambda}(1 - \epsilon_p) \int_{D_p} h(\mathcal{Y}; \tilde{\delta}, 0) d\mathcal{Y} + \epsilon_p \int_{D_p^c} h(\mathcal{Y}; \tilde{\delta}, \tau_p) d\mathcal{Y} \right\}.$$



Define  $W_j(\tilde{\delta}) = e'_j \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta})$ . Then under  $H_{0j}$ ,  $W_j(\tilde{\delta}) \sim N(0, n_1 \tilde{\omega}_{jj})$  since  $\alpha = 0$ , and under  $H_{1j}$ ,  $W_j(\tilde{\delta}) \sim N(n_1 \tau_p \tilde{\omega}_{jj}, n_1 \tilde{\omega}_{jj})$ . Then,

$$H(\tilde{\delta}, \tau_p) = p^{-\Lambda}(1 - \epsilon_p) + \epsilon_p - 2 \left\{ p^{-\Lambda}(1 - \epsilon_p) \bar{\Phi}\left(\frac{\lambda_p}{\sqrt{n_1 \tilde{\omega}_{jj}}}\right) + \epsilon_p \Phi\left(\frac{\lambda_p - n_1 \tau_p \tilde{\omega}_{jj}}{\sqrt{n_1 \tilde{\omega}_{jj}}}\right) \right\},$$

where  $\lambda_p = 1/\tau_p \{\log p^{-\Lambda} + \log(\frac{1-\epsilon_p}{\epsilon_p}) + \frac{n_1}{2} \tau_p^2 \tilde{\omega}_{jj}\}$ . Then combining (A.8) and (A.9), we have

$$H_j \geq p^{-\Lambda}(1 - \epsilon_p) \bar{\Phi}\left(\frac{\lambda_p}{\sqrt{n_1 \tilde{\omega}_{jj}}}\right) + \epsilon_p \Phi\left(\frac{\lambda_p - n_1 \tau_p \tilde{\omega}_{jj}}{\sqrt{n_1 \tilde{\omega}_{jj}}}\right). \quad (\text{A.10})$$

Using  $\epsilon_p = p^{-\beta}$ ,  $\tau_p^2 = 2r \log p / n$  and  $n_1 \tilde{\omega}_{jj} = n \omega_{jj}$  where  $\omega_{jj}$  is the  $j$ th diagonal element of  $\Omega$ , we have

$$\frac{\lambda_p}{\sqrt{n_1 \tilde{\omega}_{jj}}} = \left( \frac{\beta - \Lambda}{\sqrt{2r \omega_{jj}}} + \frac{\sqrt{r \omega_{jj}}}{\sqrt{2}} \right) \sqrt{\log p},$$

and

$$\frac{\lambda_p - n_1 \tau_p \tilde{\omega}_{jj}}{\sqrt{n_1 \tilde{\omega}_{jj}}} = \left( \frac{\beta - \Lambda - r \omega_{jj}}{\sqrt{2r \omega_{jj}}} \right) \sqrt{\log p}.$$

First, if  $\beta - r \omega_{jj} < \Lambda < \beta + r \omega_{jj}$ , then (A.10) becomes

$$\begin{aligned} H_j &\geq p^{-\Lambda} L_p p^{-\frac{(r \omega_{jj} + \beta - \Lambda)^2}{4r \omega_{jj}}} + p^{-\beta} L_p p^{-\frac{(r \omega_{jj} - \beta + \Lambda)^2}{4r \omega_{jj}}} \\ &= p^{-\beta} L_p p^{-\frac{(r \omega_{jj} - \beta + \Lambda)^2}{4r \omega_{jj}}}. \end{aligned} \quad (\text{A.11})$$

Next, we consider  $\Lambda < \beta - r \omega_{jj}$ , then (A.10) becomes

$$\begin{aligned} H_j &\geq p^{-\Lambda} L_p p^{-\frac{(r \omega_{jj} + \beta - \Lambda)^2}{4r \omega_{jj}}} + p^{-\beta} \\ &= p^{-\beta} \{1 + o(1)\}. \end{aligned} \quad (\text{A.12})$$

Last, if  $\Lambda > \beta + r \omega_{jj}$ , then (A.10) becomes

$$\begin{aligned} H_j &\geq p^{-\Lambda} + p^{-\beta} L_p p^{-\frac{(r \omega_{jj} - \beta + \Lambda)^2}{4r \omega_{jj}}} \\ &= p^{-\Lambda} \{1 + o(1)\}. \end{aligned} \quad (\text{A.13})$$

Recall that  $H = \sum_{j=1}^p H_j$ . Using the fact that  $\underline{\omega} \leq \omega_{jj} \leq \bar{\omega}$  and  $(r \omega_{jj} + \beta - \Lambda)^2 / (4r \omega_{jj})$  is an increasing function of  $\omega_{jj}$ , Theorem 1 can be derived based on the results given in (A.11), (A.12) and (A.13).

## A.5. Proof of Theorem 2

Recall that in the proof of Theorem 1, we have defined the loss function  $L(\theta, \hat{\theta}) = \sum_{i=1}^p \{\theta_i(1 - \hat{\theta}_i) + p^{-\Lambda}(1 - \theta_i)\hat{\theta}_i\}$ . For any decision rule  $\hat{\theta}_i$ , the marginal false discovery rate

$$\text{mFDR} = \frac{\mathbb{E}\{\sum_i (1 - \theta_i)\hat{\theta}_i\}}{\mathbb{E}(\sum_i \hat{\theta}_i)} = 1 - \frac{\mathbb{E}(\sum_i \theta_i \hat{\theta}_i)}{\mathbb{E}(\sum_i \hat{\theta}_i)}. \quad (\text{A.14})$$

Since  $\mathbb{E}(\sum_i \theta_i \hat{\theta}_i) \leq \min\{p^{1-\beta}, \mathbb{E}(\sum_i \hat{\theta}_i)\}$ ,  $\text{mFDR} = 1 + o(1)$  if  $p^{1-\beta} = o\{\mathbb{E}(\sum_i \hat{\theta}_i)\}$ . Hence, if mFDR is controlled at a level  $\alpha < 1$ , we must have either  $p^{1-\beta} \sim \mathbb{E}(\sum_i \hat{\theta}_i)$  or  $\mathbb{E}(\sum_i \hat{\theta}_i) = o(p^{1-\beta})$ . For the latter,

$$\text{mFNR} = \frac{\mathbb{E}\{\sum_i \theta_i(1 - \hat{\theta}_i)\}}{\mathbb{E}\{\sum_i (1 - \hat{\theta}_i)\}} = \frac{p^{1-\beta}\{1 + o(1)\}}{p\{1 + o(1)\}} = p^{-\beta}\{1 + o(1)\}.$$

Next, we consider the mFNR under the constraint  $p^{1-\beta} \sim \mathbb{E}(\sum_i \hat{\theta}_i)$ . Toward this end, we first note that if  $\text{mFDR} \leq \alpha < 1$ , the following result can be derived from (A.14):

$$p^{-\beta}\alpha \sum_{i=1}^p \mathbb{E}(\hat{\theta}_i | \theta_i = 1) \geq (1 - \alpha) \sum_{i=1}^p \mathbb{E}(\hat{\theta}_i | \theta_i = 0). \quad (\text{A.15})$$

Recall that in the proof of Theorem 1, the optimal decision rule

$$\hat{\theta}_i = \mathbb{I}\left\{\frac{(1 - p^\beta)f_0(\mathcal{Y})}{p^{-\beta}f_1(\mathcal{Y})} \leq p^\Lambda\right\},$$

where  $f_0(\mathcal{Y})$  and  $f_1(\mathcal{Y})$  are defined in (A.7). For simplicity, we choose point mass for  $\alpha = \sqrt{2r\log p/n}$  and  $\tilde{\delta}$  in (A.7). Then the decision rule can be simplified as

$$\hat{\theta}_i = \mathbb{I}\left\{e'_i \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta}) \geq \frac{(\beta - \Lambda)\log p}{\sqrt{2r\log p/n}} + \frac{\omega_{ii}r\log p}{\sqrt{2r\log p/n}}\right\}.$$

Since under  $H_{0i}$ ,  $e'_i \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta}) \sim N(0, n\omega_{ii})$ , and under  $H_{1i}$ ,  $e'_i \tilde{\Sigma}^{-1} \sum_{i=1}^{n_1} (Y_i - \tilde{\delta}) \sim N(\sqrt{2rn\log p}\omega_{ii}, n\omega_{ii})$ , we have

$$\mathbb{E}(\hat{\theta}_i | \theta_i = 0) = \bar{\Phi}\left\{\left(\frac{\beta - \Lambda}{\sqrt{2\omega_{ii}r}} + \sqrt{\frac{\omega_{ii}r}{2}}\right)\sqrt{\log p}\right\}, \quad (\text{A.16})$$

and

$$\mathbb{E}(\hat{\theta}_i | \theta_i = 1) = \bar{\Phi}\left\{\left(\frac{\beta - \Lambda}{\sqrt{2\omega_{ii}r}} - \sqrt{\frac{\omega_{ii}r}{2}}\right)\sqrt{\log p}\right\}, \quad (\text{A.17})$$

Since  $\beta - \underline{\omega}r < \Lambda < \beta + \underline{\omega}r$ ,

$$E(\hat{\theta}_i | \theta_i = 0) = L_p p^{-\frac{(\beta - \Lambda + \omega_{ii}r)^2}{4\omega_{ii}r}},$$

and

$$E(\hat{\theta}_i | \theta_i = 1) = 1 - L_p p^{-\frac{(\omega_{ii}r - \beta + \Lambda)^2}{4\omega_{ii}r}}.$$

First note that in both (A.16) and (A.17), if we choose  $\Lambda$  to be  $\Lambda_{ii} = \omega_{ii}r + \beta - 2\sqrt{\omega_{ii}r\beta\left(1 - \frac{g(\alpha, p)}{\beta}\right)}$  where  $g(\alpha, p) = \log\left\{\frac{\alpha}{(1-\alpha)}\sqrt{4\pi\beta\log p}\right\}\log^{-1}p$ , then the “=” holds in (A.15). To have a universal  $\Lambda$  which does not depend on index  $i$ , we can choose  $\Lambda = \underline{\omega}r + \beta - 2\sqrt{\underline{\omega}r\beta\left(1 - \frac{g(\alpha, p)}{\beta}\right)}$  such that the right hand is no greater than the left hand of (A.15). Equivalently, this implies that  $\text{mFDR} \leq \alpha < 1$ .

Given  $\Lambda$ ,

$$\begin{aligned} \text{mFNR} &= \frac{E\{\sum_i \theta_i(1 - \hat{\theta}_i)\}}{E\{\sum_i (1 - \hat{\theta}_i)\}} = \frac{\sum_i L_p p^{-\beta - \frac{(\omega_{ii}r - \beta + \Lambda)^2}{4\omega_{ii}r}}}{p\{1 + o(1)\}} \\ &\geq \frac{L_p p^{1 - \beta - \frac{(\bar{\omega}r - \beta + \Lambda)^2}{4\bar{\omega}r}}}{p\{1 + o(1)\}} \\ &\geq L_p p^{-\beta - \left\{\sqrt{\bar{\omega}r} - \sqrt{\beta - g(\alpha, p)}\right\}^2}. \end{aligned}$$

This completes the proof of Theorem 2.

## A.6. Proof of Theorem 3

Note that the loss function  $L\{\theta_j, \text{sgn}(\hat{\delta}_j)\} = \theta_j\{1 - \text{sgn}(\hat{\delta}_j)\} + p^{-\Lambda}(1 - \theta_j)\text{sgn}(\hat{\delta}_j)$  where  $\theta_j = 0$  if  $\delta_j = 0$  and  $\theta_j = 1$  otherwise, and  $\hat{\delta}_j$  is estimated to be one of three values from  $\{-\delta^{data}, 0, \delta^{date}\}$  by the DATE procedure. Since after the thresholding step, all the coordinates are assigned into either  $\mathcal{U}(s)$  or  $\mathcal{U}^c(s)$ , the corresponding risk is

$$H(\Lambda) = \sum_{j=1}^p E(L\{\theta_j, \text{sgn}(\hat{\delta}_j)\}) = I + II,$$

where  $I$  is the risk in the thresholding step and  $II$  is the risk in the excising step, i.e.,

$$I = \sum_{j=1}^p E(L\{\theta_j, \text{sgn}(\hat{\delta}_j)\}I\{j \notin \mathcal{U}(s)\}), \quad II = \sum_{j=1}^p E(L\{\theta_j, \text{sgn}(\hat{\delta}_j)\}I\{j \in \mathcal{U}(s)\}).$$

For  $I$ , we know that if  $j \notin \mathcal{U}(s)$ , the estimated signal  $\hat{\delta}_j = 0$  based on the DATE procedure. By Lemma 3,

$$I = \sum_{j=1}^p \mathbb{P}(\hat{T}_j < 2s \log p, \delta_j \neq 0) \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)},$$

which is not greater than the upper bound of  $H(\Lambda)$  except a slowly varying function. Hence, we only need to show that  $II \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)}$ .

Let event  $A_p = \{\mathcal{U}(s)\}$  are split into disconnected clusters of size no more than  $K$  with respect to  $(V_0, \Omega^*)$ . By Lemma 4,  $\mathbb{P}(A_p^c) \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)}$ . Therefore, it is sufficient to show that for all  $1 \leq j \leq p$ ,

$$\mathbb{E} \left( L\{\theta_j, \text{sgn}(\hat{\delta}_j)\} \mathbb{I}\{(j \in \mathcal{U}(s)) \cap A_p\} \right) \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)}.$$

By Lemma 4, we know that over the event  $\{j \in \mathcal{U}(s)\} \cap A_p$ , there exists a unique component  $\mathcal{I}_0 = \{i_1, \dots, i_m\}$  with size  $m \leq K$  satisfying  $j \in \mathcal{I}_0$ . Therefore, it is sufficient to show that for any fixed connected subgroup  $\mathcal{I}_0$  that contains  $j$ ,

$$\mathbb{E} \left( L\{\theta_j, \text{sgn}(\hat{\delta}_j)\} \mathbb{I}\{(j \in \mathcal{I}_0) \cap A_p\} \right) \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)}. \quad (\text{A.18})$$

Recall that  $L\{\theta_j, \text{sgn}(\hat{\delta}_j)\}$  consists of the expected false positive and false negative. Accordingly if we define two events

$$R_1 = \{\theta_j = 0, \text{sgn}(\hat{\delta}_j) \neq 0, j \in \mathcal{I}_0, A_p\},$$

and

$$R_2 = \{\theta_j = 1, \text{sgn}(\hat{\delta}_j) = 0, j \in \mathcal{I}_0, A_p\},$$

then to show (A.18), we only need to show that

$$p^{-\Lambda} \mathbb{P}(R_1) \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)}, \quad \mathbb{P}(R_2) \leq L_p p^{1 - \{(\omega r - \beta + \Lambda)^2 + 4\omega r \beta\} / (4\omega r)}. \quad (\text{A.19})$$

Within the component  $\mathcal{I}_0$ , by conducting the DATE procedure, it is possible that some signals are wrongly identified as noise and some noise can be identified

as signals. For convenience, we let  $B_{nn}$  be the number of true negatives,  $B_{ns}$  be the number of false positives,  $B_{sn}$  be the number of false negatives, and  $B_{ss}$  be the number of true positives. Then, the total number of signals in  $\mathcal{I}_0$  is  $B_{sn} + B_{ss}$ . Let the event  $M_p = \{\text{sgn}(\hat{\delta}(\mathcal{I}_0) \neq \text{sgn}(\delta^{\mathcal{I}_0}), j \in \mathcal{I}_0, A_p\}$ . Since  $j \in \mathcal{I}_0$ , the event  $R_1$  is contained in  $M_p$  and the event  $R_2$  is contained in the event  $M_p \cap \{B_{sn} + B_{ss} \geq 1\}$ . Therefore, to show (A.19), we only need to show

$$\begin{aligned} p^{-\Lambda} \mathbb{P}(M_p) &\leq L_p p^{-\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)}, \\ \mathbb{P}(M_p \cap \{B_{sn} + B_{ss} \geq 1\}) &\leq L_p p^{-\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)}. \end{aligned} \quad (\text{A.20})$$

Let the event  $D_p = \{\|\hat{\Omega} - \Omega\|_{L_1} \leq (\log p/n)^{(1-\alpha)/2}\}$ . Since  $\mathbb{P}(D_p^c) = o(p^{-1})$ , it is sufficient to show (A.20) over the event  $D_p$ . Moreover, define the event  $B_p(\mathcal{I}_0)$  through its complement:  $B_p^c(\mathcal{I}_0) = \{\text{there exist indices } i \notin \mathcal{I}_0 \text{ and } j \in \mathcal{I}_0 \text{ such that } \delta_i \neq 0, \Omega^*(i, j) \neq 0\}$ . Similar to Ji and Jin (2012), we can show that

$$\mathbb{P}(j \in \mathcal{I}_0, B_p^c \cap A_p) \leq L_p p^{-\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)}.$$

Therefore, in order to show (A.20), it is sufficient to show that

$$\begin{aligned} p^{-\Lambda} \mathbb{P}(M_p \cap B_p \cap D_p) &\leq L_p p^{-\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)}, \\ \mathbb{P}(M_p \cap B_p \cap D_p \cap \{B_{sn} + B_{ss} \geq 1\}) &\leq L_p p^{-\{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r)}. \end{aligned} \quad (\text{A.21})$$

Most importantly, by Lemma A.4 of Ji and Jin (2012), over the event  $\{(j \in \mathcal{I}_0) \cap A_p \cap B_p\}$ ,

$$\|(\Omega\delta)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0\mathcal{I}_0}\delta^{\mathcal{I}_0}\|_{\infty} = o(\sqrt{\log p/n}),$$

which implies that  $(\Omega\delta)^{\mathcal{I}_0} \approx \Omega^{\mathcal{I}_0\mathcal{I}_0}\delta^{\mathcal{I}_0}$ . This enables us to find  $\hat{\delta}(\mathcal{I}_0)$  over the event  $M_p \cap B_p \cap D_p$ , each components of which has the value taken from  $\{-\delta^{\text{date}}, 0, \delta^{\text{date}}\}$  to minimize

$$n\{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} - A\delta\}' A^{-1}\{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} - A\delta\} + (\lambda^{\text{date}})^2 \|\delta\|_0, \quad (\text{A.22})$$

where  $A = \Omega^{\mathcal{I}_0 \mathcal{I}_0}$ .

If the event  $\{M_p \cap B_p \cap D_p\}$  happens, then by (A.22),

$$\begin{aligned} & n\{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0 \mathcal{I}_0} \delta(\mathcal{I}_0)\}' (\Omega^{\mathcal{I}_0 \mathcal{I}_0})^{-1} \{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0 \mathcal{I}_0} \hat{\delta}(\mathcal{I}_0)\} + (\lambda^{date})^2 \|\hat{\delta}(\mathcal{I}_0)\|_0 \\ \leq & n\{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0 \mathcal{I}_0} \delta^*(\mathcal{I}_0)\}' (\Omega^{\mathcal{I}_0 \mathcal{I}_0})^{-1} \{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} - \Omega^{\mathcal{I}_0 \mathcal{I}_0} \delta^*(\mathcal{I}_0)\} + (\lambda^{date})^2 \|\delta^*(\mathcal{I}_0)\|_0, \end{aligned}$$

where  $\delta^*(\mathcal{I}_0)$  is defined to be a vector on  $\mathcal{I}_0$  each component of which corresponds to the true signals or noise in the sense that for  $l \in \mathcal{I}_0$ ,  $\delta_l^* = 0$  if  $\delta_l = 0$  and  $\delta_l^* = \delta^{date} \text{sgn}(\delta)$  if  $\delta_l \neq 0$ . If we let  $d = \|\delta^*(\mathcal{I}_0)\|_0 - \|\hat{\delta}(\mathcal{I}_0)\|_0 = B_{sn} - B_{ns}$ , it follows that

$$n\{\delta^*(\mathcal{I}_0) - \hat{\delta}(\mathcal{I}_0)\}' (\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0} \leq \frac{1}{2} \left( (\lambda^{date})^2 d + n\{\delta^*(\mathcal{I}_0)\}' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \delta^*(\mathcal{I}_0) - \{\hat{\delta}(\mathcal{I}_0)\}' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \hat{\delta}(\mathcal{I}_0) \right). \quad (\text{A.23})$$

Since  $\sqrt{n}\{(\bar{Z}_1 - \bar{Z}_2)^{\mathcal{I}_0}\} = \sqrt{n} \Omega^{\mathcal{I}_0 \mathcal{I}_0} \delta^{\mathcal{I}_0} + z$  where  $z \sim N(0, \Omega^{\mathcal{I}_0 \mathcal{I}_0})$ . Then, (A.23) can be written as

$$\frac{\Delta_1' z}{\sqrt{\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \leq -\frac{\sqrt{2r \log p}}{2\sqrt{\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \left\{ -d \frac{\beta - \Lambda}{r} + 2\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1 \right\}, \quad (\text{A.24})$$

where

$$\Delta_1 = \frac{\sqrt{n}\{(\delta^*)^{\mathcal{I}_0} - \hat{\delta}(\mathcal{I}_0)\}}{\sqrt{2r \log p}}, \quad \Delta_2 = \frac{\sqrt{n}\{\delta^{\mathcal{I}_0} - (\delta^*)^{\mathcal{I}_0}\}}{\sqrt{2r \log p}}.$$

In (A.24), both  $\delta^{\mathcal{I}_0}$  and  $z$  are random. Given  $\delta^{\mathcal{I}_0}$ ,

$$\frac{\Delta_1' z}{\sqrt{\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \sim N(0, 1).$$

Note that if the event  $\{M_p \cap B_p \cap D_p\}$  happens, then the inequality (A.24) holds.

Therefore,

$$\begin{aligned} & P(M_p \cap B_p \cap D_p) \\ \leq & P\left( \frac{\Delta_1' z}{\sqrt{\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \leq -\frac{\sqrt{2r \log p}}{2\sqrt{\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \left\{ -d \frac{\beta - \Lambda}{r} + 2\Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta_1' \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1 \right\} \right). \end{aligned} \quad (\text{A.25})$$

Next, we evaluate the right hand side of the inequality (A.25) for different values of  $B_{sn} + B_{ss}$ . To this end, we first notice that the right hand side is bounded by  $p^{-\beta(B_{sn}+B_{ss})}$ , which is the probability of having  $B_{sn} + B_{ss}$  signals in  $\mathcal{I}_0$ . Therefore, if  $B_{sn} + B_{ss} \geq \{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r\beta)$ , from (A.25), it is easy to show that (A.21) is true. Hence, we only need to consider the case where  $B_{sn} + B_{ss} < \{(\underline{\omega}r - \beta + \Lambda)^2 + 4\underline{\omega}r\beta\}/(4\underline{\omega}r\beta)$ . Note that the value of  $B_{nn}$  does not affect the inequality in (A.25). Therefore, we assume  $B_{nn} = 0$ . Also similar to Lemma A.6 of Ji and Jin (2012), it can be shown that  $\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1 \geq \underline{\omega}$ . Moreover, since the support of any signal  $|\delta_k|$  is  $[\sqrt{2r \log p/n}, (1 + \eta)\sqrt{2r \log p/n}]$ ,  $\Delta_2 \geq 0$  if  $\text{sgn}(\delta_k) = 1$  and  $\Delta_2 < 0$  otherwise.

- $B_{sn} + B_{ss} = 0$ ;

For this case, we have  $d = B_{sn} - B_{ns} = -B_{ns} \leq -1$ . Using the fact that

$\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1 \geq \underline{\omega}$  and  $\Delta_2 = 0$ , we have

$$\frac{-d \frac{\beta - \Lambda}{r} + 2\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \geq \frac{\frac{\beta - \Lambda}{r} + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \geq \frac{\frac{\beta - \Lambda}{r} + \underline{\omega}}{2\sqrt{\underline{\omega}}}.$$

Then, from (A.25), by using the fact that  $\bar{\Phi}(a) \approx \phi(a)/a$  if  $a \rightarrow \infty$ , we have

$$\mathbb{P}(M_p \cap B_p \cap D_p) \leq \bar{\Phi}\left(-\frac{\frac{\beta - \Lambda}{r} + \underline{\omega}}{2\sqrt{\underline{\omega}}}\sqrt{2r \log p}\right) \leq L_p p^{-\frac{(\omega r + \beta - \Lambda)^2}{4\omega r}}.$$

Then, for this case,

$$\mathbb{P}(R_1) \leq L_p p^{-\frac{(\omega r + \beta - \Lambda)^2}{4\omega r}}. \quad (\text{A.26})$$

- $B_{sn} + B_{ss} = 1$  but  $B_{ns} = 0$ ;

For this case, since  $\text{sgn}(\hat{\delta}(\mathcal{I}_0)) \neq \text{sgn}(\delta^{\mathcal{I}_0})$ , we must have  $B_{sn} \neq 0$ . Otherwise, both  $B_{sn} = 0$  and  $B_{ns} = 0$  leads to  $\text{sgn}(\hat{\delta}(\mathcal{I}_0)) = \text{sgn}(\delta^{\mathcal{I}_0})$ . As a result,  $B_{sn} = 1$  and  $B_{ss} = 0$ . It follows that  $d = B_{sn} - B_{ns} = 1$ ,  $\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1 \geq \underline{\omega}$ , and  $\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 \geq 0$ . Then,

$$\frac{-d \frac{\beta - \Lambda}{r} + 2\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \geq \frac{-\frac{\beta - \Lambda}{r} + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \geq \frac{-\frac{\beta - \Lambda}{r} + \underline{\omega}}{2\sqrt{\underline{\omega}}}.$$

This, together with  $B_{sn} + B_{ss} = 1$ , shows that (A.25) satisfies

$$P(M_p \cap B_p \cap D_p) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}},$$

which implies that

$$P(R_1) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}}, P(R_2) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}}. \quad (\text{A.27})$$

In the following, we consider  $1 \leq B_{sn} + B_{ss} \leq \{(\omega r - \beta + \Lambda)^2 + 4\omega r\beta\}/(4\omega r\beta)$ ,  $B_{nn} = 0$ , and when  $B_{ns} = 0$ ,  $B_{sn} + B_{ss} \geq 2$ . To this end, we apply the Cauchy-Schwartz to get

$$|\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2| \leq \sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1} \sqrt{\Delta'_2 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2}. \quad (\text{A.28})$$

Using the spectral decomposition,  $\Delta'_2 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 \leq C_0 \|\Delta_2\|_2^2$  where  $C_0$  is defined in condition (C2). Since the support of signal is  $[\sqrt{2s \log p/n}, (1+\eta)\sqrt{2s \log p/n}]$ , and  $\Delta_2$  has  $(B_{ss} + B_{sn})$  nonzero signals,

$$\Delta'_2 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 \leq C_0 (B_{ss} + B_{sn}) \eta^2.$$

Moreover, with assumption  $B_{sn} + B_{ss} \leq \{(\omega r - \beta + \Lambda)^2 + 4\omega r\beta\}/(4\omega r\beta)$ , (A.28) can be written as

$$|\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2| \leq \sqrt{C} \cdot \sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1},$$

where  $C = C_0 \eta^2 \{(\omega r - \beta + \Lambda)^2 + 4\omega r\beta\}/(4\omega r\beta)$ . Then it follows that  $\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 \geq -\sqrt{C} \cdot \sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}$ , which implies that

$$\frac{-d \frac{\beta - \Lambda}{r} + 2\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \geq \frac{-d \frac{\beta - \Lambda}{r} + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} - \sqrt{C}.$$

Next we consider three different cases:  $B_{ns} = B_{sn} \geq 1$ ;  $B_{ns} > B_{sn}$ ;  $B_{ns} < B_{sn}$ .

For the case  $B_{ns} = B_{sn} \geq 1$ , we have  $d = B_{sn} - B_{ns} = 0$ . Then,

$$\frac{-d \frac{\beta - \Lambda}{r} + 2\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \geq \frac{\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}}{2} - \sqrt{C}.$$



Recall that  $\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1 \geq \underline{\omega}$  and  $\eta \leq \frac{\beta - \Lambda}{\sqrt{C_0 r}} \frac{\sqrt{\beta r}}{\sqrt{(\underline{\omega} r - \beta + \Lambda)^2 + 4\underline{\omega} r \beta}}$ . Then,  $\sqrt{C} \leq (\beta - \Lambda)/(2\sqrt{\underline{\omega} r})$ , and

$$P(M_p \cap B_p \cap D_p) \leq L_p p^{-\beta} p^{-\frac{(\underline{\omega} r - \beta + \Lambda)^2}{4\underline{\omega} r}}.$$

For the case  $B_{ns} > B_{sn}$ ,  $d = B_{sn} - B_{ns} \leq -1$ . Since  $\eta \leq \frac{\beta - \Lambda}{\sqrt{C_0 r}} \frac{\sqrt{\beta r}}{\sqrt{(\underline{\omega} r - \beta + \Lambda)^2 + 4\underline{\omega} r \beta}}$ , we have  $\sqrt{C} \leq (\beta - \Lambda)/(\sqrt{\underline{\omega} r})$ , and

$$\begin{aligned} \frac{-d \frac{\beta - \Lambda}{r} + 2\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} &\geq \frac{\frac{\beta - \Lambda}{r} + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} - \sqrt{C} \\ &\geq \frac{1}{2} \left( \frac{\beta - \Lambda}{\sqrt{\underline{\omega} r}} + \sqrt{\underline{\omega}} \right) - \sqrt{C} \\ &\geq \frac{1}{2} \left( \sqrt{\underline{\omega}} - \frac{\beta - \Lambda}{\sqrt{\underline{\omega} r}} \right). \end{aligned}$$

Since  $B_{sn} + B_{ss} \geq 1$ , then

$$P(M_p \cap B_p \cap D_p) \leq L_p p^{-\beta} p^{-\frac{(\underline{\omega} r - \beta + \Lambda)^2}{4\underline{\omega} r}}.$$

For the case  $B_{ns} < B_{sn}$ , we have either  $B_{ns} = 0$  or  $B_{ns} \geq 1$ . If  $B_{ns} = 0$ ,  $B_{sn} + B_{ss} \geq 2$  as we have required. If  $B_{ns} \geq 1$ , we also have  $B_{sn} + B_{ss} \geq 2$  due to the fact that  $B_{sn} > B_{ns}$ . Since  $\beta + (\underline{\omega} r - \beta + \Lambda)^2 / (4\underline{\omega} r) = \Lambda + (\underline{\omega} r + \beta - \Lambda)^2 / (4\underline{\omega} r)$  and  $\beta - \Lambda < \underline{\omega} r$ , then

$$\beta + (\underline{\omega} r - \beta + \Lambda)^2 / (4\underline{\omega} r) \leq \Lambda + \underline{\omega} r.$$

Then using the fact that  $\beta(B_{sn} + B_{ss}) \geq 2\beta$  and by assuming  $\beta > 1/2$  and  $\underline{\omega} r < (\sqrt{1 - \Lambda} + \sqrt{1 - \beta})^2$ , we can derive the following inequality

$$\beta(B_{sn} + B_{ss}) \geq \beta + \frac{(\underline{\omega} r - \beta + \Lambda)^2}{4\underline{\omega} r}.$$

Therefore, for three different cases:  $B_{ns} = B_{sn} \geq 1$ ;  $B_{ns} > B_{sn}$ ;  $B_{ns} < B_{sn}$ , we have

$$P(M_p \cap B_p \cap D_p) \leq L_p p^{-\beta} p^{-\frac{(\underline{\omega} r - \beta + \Lambda)^2}{4\underline{\omega} r}},$$

which implies that

$$P(R_1) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}}, P(R_2) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}}. \quad (\text{A.29})$$

In summary, from (A.26), (A.27) and (A.29), we know that if  $B_{sn} + B_{ss} = 0$ ,  $P(R_1) \leq L_p p^{-\frac{(\omega r + \beta - \Lambda)^2}{4\omega r}}$ . And if  $B_{sn} + B_{ss} \geq 1$ ,  $P(R_1) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}}$ . Since

$$\frac{(\omega r + \beta - \Lambda)^2}{4\omega r} \leq \beta + \frac{(\omega r - \beta + \Lambda)^2}{4\omega r},$$

we have  $P(R_1) \leq L_p p^{-\frac{(\omega r + \beta - \Lambda)^2}{4\omega r}}$ . Similarly, we have  $P(R_2) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}}$ . Then,

$$p^{-\Lambda} P(R_1) + P(R_2) \leq L_p p^{-\beta} p^{-\frac{(\omega r - \beta + \Lambda)^2}{4\omega r}},$$

which shows that (A.19) is true. This completes the proof of Theorem 3.

## A.7. Proof of Theorem 4

Let  $\theta_i = 0$  if  $\delta_i = 0$  and  $\theta_i = 1$  otherwise, and  $\hat{\delta}_i$  is the corresponding estimate by the DATE procedure. Recall that the marginal false discovery rate is defined as

$$\text{mFDR} = \frac{\sum_i P(\theta_i = 0, \text{sgn}(\hat{\delta}_i) \neq 0)}{\sum_i P(\theta_i = 0, \text{sgn}(\hat{\delta}_i) \neq 0) + \sum_i P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) \neq 0)}.$$

Similar to (A.25) in the proof of Theorem 3,

$$P(\theta_i \neq \text{sgn}(\hat{\delta}_i)) \leq p^{-\beta(B_{sn} + B_{ss})} \bar{\Phi} \left( \frac{-d \frac{\beta - \Lambda - \frac{\Upsilon}{2 \log p}}{r} + 2\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_2 + \Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}{2\sqrt{\Delta'_1 \Omega^{\mathcal{I}_0 \mathcal{I}_0} \Delta_1}} \sqrt{2r \log p} \right). \quad (\text{A.30})$$

First, note that

$$\begin{aligned} P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) \neq 0) &= P(\theta_i = 1) - P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) = 0) \\ &= p^{-\beta} - P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) = 0). \end{aligned}$$

Following the similar derivations for Theorem 3, from (A.30), we can show that

$$P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) = 0) \leq p^{-\beta} \bar{\Phi} \left( \frac{(-\beta + \Lambda + \omega r) \log p + \Upsilon/2}{\sqrt{2\omega r \log p}} \right),$$

where  $\Upsilon$  is defined in Theorem 4. Then,

$$P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) \neq 0) \geq p^{-\beta} \bar{\Phi} \left( \frac{(\beta - \Lambda - \underline{\omega}r) \log p - \Upsilon/2}{\sqrt{2\underline{\omega}r \log p}} \right).$$

Similarly,

$$\begin{aligned} P(\theta_i = 0, \text{sgn}(\hat{\delta}_i) \neq 0) &= P(\theta_i = 0, \text{sgn}(\hat{\delta}_i) = 1) + P(\theta_i = 0, \text{sgn}(\hat{\delta}_i) = -1) \\ &\leq 2(1 - p^{-\beta}) \bar{\Phi} \left( \frac{(\beta - \Lambda + \underline{\omega}r) \log p - \Upsilon/2}{\sqrt{2\underline{\omega}r \log p}} \right). \end{aligned}$$

To require mFDR to be controlled at  $\alpha$ , we need to find  $\Upsilon$  such that

$$\alpha \geq \frac{2(p - p^{1-\beta}) \bar{\Phi} \left( \frac{(\beta - \Lambda + \underline{\omega}r) \log p - \Upsilon/2}{\sqrt{2\underline{\omega}r \log p}} \right)}{2(p - p^{1-\beta}) \bar{\Phi} \left( \frac{(\beta - \Lambda + \underline{\omega}r) \log p - \Upsilon/2}{\sqrt{2\underline{\omega}r \log p}} \right) + p^{1-\beta} \bar{\Phi} \left( \frac{(\beta - \Lambda - \underline{\omega}r) \log p - \Upsilon/2}{\sqrt{2\underline{\omega}r \log p}} \right)}. \quad (\text{A.31})$$

When  $p \rightarrow \infty$ , (A.31) can be solved asymptotically. If we assume  $\Upsilon = o(\log p)$ , then by the fact that  $\underline{\omega}r > \beta - \Lambda$ ,

$$\bar{\Phi} \left( \frac{(\beta - \Lambda - \underline{\omega}r) \log p - \Upsilon/2}{\sqrt{2\underline{\omega}r \log p}} \right) \rightarrow 1.$$

Then using the fact that  $\bar{\Phi}(a) \approx \phi(a)/a$  for  $a \rightarrow \infty$ , we can solve (A.31) by choosing

$$\Upsilon = \frac{4\underline{\omega}r}{\underline{\omega}r + \beta - \Lambda} \left( \frac{1}{2} \log \log p + \log \left\{ \frac{\alpha \sqrt{\pi} (\underline{\omega}r + \beta - \Lambda)}{2 \sqrt{\underline{\omega}r} (1 - \alpha)} \right\} \right)$$

such that

$$\begin{aligned} \text{mFDR} &\leq \frac{\frac{\alpha}{1-\alpha} p^{-\frac{(\underline{\omega}r + \beta - \Lambda)^2}{4\underline{\omega}r}}}{\frac{\alpha}{1-\alpha} p^{-\frac{(\underline{\omega}r + \beta - \Lambda)^2}{4\underline{\omega}r}} + p^{-\beta} \{1 + o(1)\}} \\ &\leq \alpha \{1 + o(1)\}, \end{aligned}$$

where we have used the result that  $\Lambda = (\sqrt{\underline{\omega}r} - \sqrt{\beta})^2$ .

Similarly, the marginal false non-discovery rate

$$\begin{aligned} \text{mFNR} &= \frac{\sum_i P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) = 0)}{\sum_i P(\theta_i = 1, \text{sgn}(\hat{\delta}_i) = 0) + \sum_i P(\theta_i = 0, \text{sgn}(\hat{\delta}_i) = 0)} \\ &\leq L_p p^{-\beta - (\sqrt{\underline{\omega}r} - \sqrt{\beta})^2}. \end{aligned}$$

This completes the proof of Theorem 4.

## A.8. Proof of Theorem 5

Note that the optimal rate does not change if we add a slowly varying logarithm function to the tuning parameters in Theorem 4 by carefully reviewing its proof. Therefore, if we can show that there exists a constant  $C$  such that

$$P(|\hat{\beta} - \beta| > (\log p)^{-C}) = o(p^{-1}), \quad (\text{A.32})$$

$$P(|\hat{r} - r| > (\log p)^{-C}) = o(p^{-1}), \quad (\text{A.33})$$

and

$$P(|\hat{\omega} - \omega| > (\log p)^{-C}) = o(p^{-1}), \quad (\text{A.34})$$

where  $\hat{\beta}$ ,  $\hat{r}$  and  $\hat{\omega}$  are defined in (4.5), then Theorem 5 can be proved.

First, let's prove (A.34). Note that with probability  $1 - O(p^{-\tau})$  and for some constant  $C$ ,

$$\|\hat{\Omega} - \Omega\|_{L_1} \leq C \left\{ \left( \frac{\log p}{n} \right)^{\frac{1-\zeta}{2}} \right\}.$$

Under condition (C4),

$$\left( \frac{\log p}{n} \right)^{\frac{1-\zeta}{2}} = (\log p)^{-\frac{(1-\theta)(1-\zeta)}{2\theta}} < (\log p)^{-1/2}.$$

If  $\tau > 1$ , for large enough  $p$ ,

$$\begin{aligned} P(|\hat{\omega} - \omega| > (\log p)^{-C}) &\leq P\left(\min_{1 \leq k \leq p} |\hat{\omega}_{kk} - \omega_{kk}| > (\log p)^{-C}\right) \\ &\leq P(\|\hat{\Omega} - \Omega\|_{L_1} > (\log p)^{-C}) = o(p^{-1}). \end{aligned} \quad (\text{A.35})$$

Next, let's prove (A.32) or equivalently, we need to show that

$$P\left(\left| \frac{1}{p^{1-\beta}} \sum_{k=1}^p \mathbf{I}(\hat{T}_k > 2q \log p) - 1 \right| > L_p p^{-C}\right) = o(p^{-1}).$$

To this end, we first notice that

$$\sum_{k=1}^p \mathbf{I}(\hat{T}_k > 2q \log p) = \sum_{k=1}^p \mathbf{I}(\hat{T}_k > 2q \log p) \mathbf{I}(\delta_k = 0) + \sum_{k=1}^p \mathbf{I}(\hat{T}_k > 2q \log p) \mathbf{I}(\delta_k \neq 0).$$

Then,

$$\begin{aligned}
& \mathbb{P}\left(\left|\frac{1}{p^{1-\beta}} \sum_{k=1}^p \mathbb{I}(\hat{T}_k > 2q \log p) - 1\right| > L_p p^{-C}\right) \\
& \leq \mathbb{P}\left(\left|\frac{1}{p^{1-\beta}} \sum_{k=1}^p \mathbb{I}(\hat{T}_k > 2q \log p) \mathbb{I}(\delta_k = 0)\right| > L_p p^{-C}\right) \\
& + \mathbb{P}\left(\left|\frac{1}{p^{1-\beta}} \sum_{k=1}^p \mathbb{I}(\hat{T}_k > 2q \log p) \mathbb{I}(\delta_k \neq 0) - 1\right| > L_p p^{-C}\right).
\end{aligned}$$

Using Chebyshev's inequality and the result in Lemma A.7 of Ji and Jin (2012), we have

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{p^{1-\beta}} \sum_{k=1}^p \mathbb{I}(\hat{T}_k > 2q \log p) \mathbb{I}(\delta_k = 0)\right| > L_p p^{-C}\right) & \leq \frac{\mathbb{E}(\{\sum_{k=1}^p \mathbb{I}(\hat{T}_k > 2q \log p) \mathbb{I}(\delta_k = 0)\}^m)}{p^{m-m\beta} L_p^m p^{-mC}} \\
& \leq \frac{L_p p^m p^{-mq}}{p^{m-m\beta-mC}} \\
& = L_p p^{-m(q-\beta+C)}.
\end{aligned}$$

Then we can choose  $m$  large enough to have  $p^{-m(q-\beta+C)} = o(p^{-1})$  since  $q > \beta$ . Based on similar derivations, (A.33) can be shown accordingly. This completes the proof of Theorem 5.

## REFERENCE

- BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, **57** 289-300.
- BENJAMINI, Y. AND YEKUTIELI, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, **29** 1165-1188.
- BICKEL, P. AND LEVINA, E. (2008a). Regularized estimation of large covariance matrices. *The Annals of Statistics*, **36**, 199-227.
- BICKEL, P. AND LEVINA, E. (2008b). Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577-2604.

- CAI, T. , LIU, W. AND XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B*, **76**, 349-372.
- CAI, T. , LIU, W. AND LUO, X. (2011). A constrained  $l_1$  minimization approach to sparse precision matrix estimation. *Journal of American Statistical Association*, **106**, 594-607.
- DONOHO, D. AND JIN, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, **32** 962-994.
- EFRON, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of American Statistical Association*, **102** 93-103.
- FRIEDMAN, J. , HASTIE, T. AND TIBSHIRANI R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9** 432-441.
- FRIEZE, A. AND MOLLOY, M. (1999). Splitting an expander graph. *Journal of Algorithms*, **33**, 166-172.
- GENOVESE, C. AND WASSERMAN, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B*, **64** 499-517.
- HALL, P. AND JIN, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, **38**, 1686-1732.
- JI, P. AND JIN, J. (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics*, **40**, 73-103.
- JI, P. AND ZHAO, Z. (2014). Rate optimal multiple testing procedure in high-dimensional regression. *Manuscript*.
- KLAUS, B. AND STRIMMER, K. (2013). Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics*, **14**, 129-143.
- QIU, X., KLEBANOV, L. AND YAKOVLEV, A. (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding

- differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, **4**, 34.
- RICHARDSON, A., WANG, Z., NICOLO, A., LU, X., BROWN, M., MIRON, A., LIAO, X., IGLEHART, J., LIVINGSTON, D. AND GANESAN, S. (2006). X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, **9**, 121-132.
- SUN, W. AND CAI, T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of American Statistical Association*, **102**, 901-912.
- SUN, W. AND CAI, T. (2009). Large-scale multiple testing under dependency. *Journal of the Royal Statistical Society: Series B*, **71**, 393-424.
- XIE, J., CAI, T. AND LI, H. (2011). Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, **98**, 273-290.
- XIE, J., CAI, T., MARIS, J. AND LI, H. (2011). Optimal false discovery rate control for dependent data. *Statistics and Its Interface*, **4**, 417-430.
- YUAN, M. AND LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19-35.
- ZHAO, T. , LIU, H. , ROEDER, K. , LAFFERTY, J. AND WASSERMAN, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, **13**, 1059-1062.

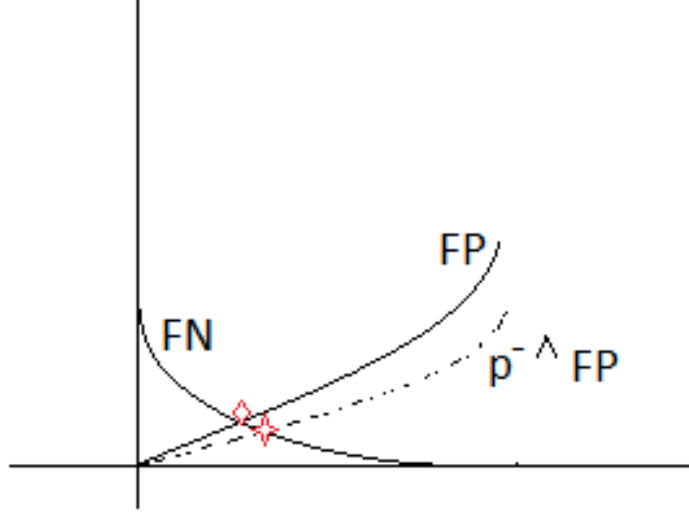


Figure 1: The horizontal axis represents the number of  $\hat{\delta}_k \neq 0$ . The diamond is the intersection point of the false positives line (FP) and the false negatives line (FN) where  $H(0)$  is minimized and the star is the intersection point where  $H(\Lambda)$  is minimized.

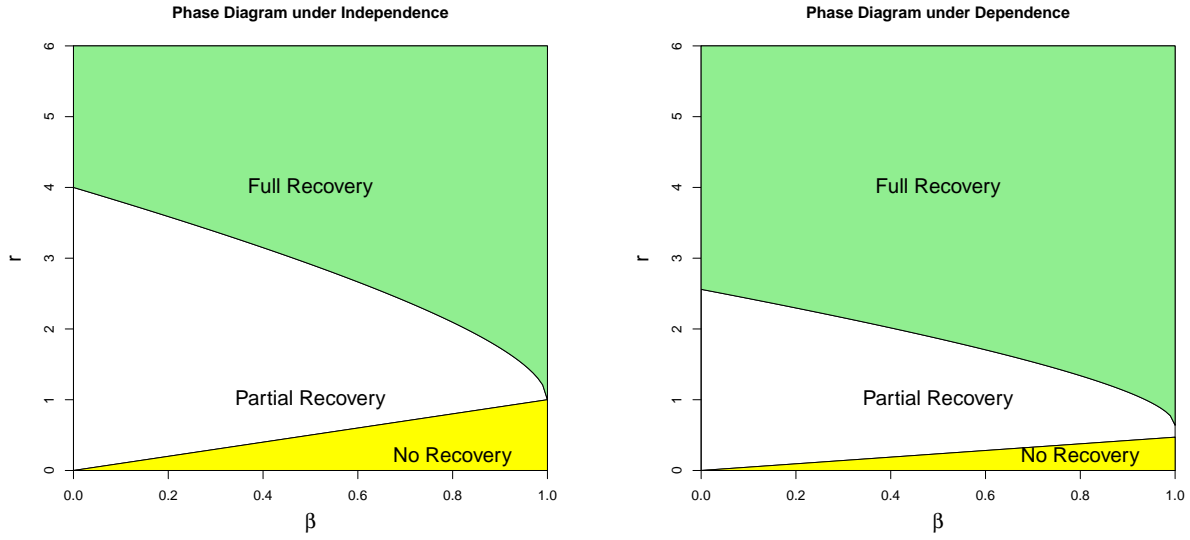


Figure 2: Left: phase diagram for signal recovery without data dependence. Right: phase diagram for signal recovery with  $\Sigma_1 = \Sigma_2 = (0.6^{|i-j|})$  for  $1 \leq i, j \leq p$ .



Table 1: The performance of  $\text{DATE}_\Omega$  and  $\text{DATE}_{\hat{\Omega}}$  in terms of mFDR and mFNR subject to different values of  $s$  and  $q$  chosen from two intervals separated by  $\beta = 0.6$  (mFDR and mFNR of  $\text{DATE}_{\hat{\Omega}}$  are included in parenthesis).

$s \backslash q$	0.65	0.70	0.75	0.80	0.85	0.90
	mFDR					
0.25	0.045(0.041)	0.047(0.042)	0.038(0.033)	0.053(0.046)	0.036(0.029)	0.038(0.030)
0.30	0.038(0.039)	0.041(0.034)	0.042(0.032)	0.043(0.030)	0.050(0.035)	0.032(0.020)
0.35	0.041(0.037)	0.033(0.040)	0.025(0.025)	0.048(0.041)	0.033(0.028)	0.048(0.041)
0.40	0.046(0.051)	0.039(0.034)	0.034(0.033)	0.031(0.034)	0.044(0.041)	0.043(0.035)
0.45	0.040(0.044)	0.036(0.039)	0.041(0.034)	0.043(0.041)	0.050(0.037)	0.037(0.025)
0.50	0.041(0.042)	0.031(0.030)	0.038(0.033)	0.042(0.033)	0.039(0.030)	0.043(0.031)
	mFNR					
0.25	0.005(0.006)	0.005(0.006)	0.006(0.007)	0.005(0.007)	0.006(0.007)	0.006(0.007)
0.30	0.006(0.007)	0.005(0.006)	0.006(0.007)	0.005(0.006)	0.005(0.006)	0.005(0.007)
0.35	0.005(0.006)	0.006(0.007)	0.005(0.006)	0.006(0.007)	0.005(0.007)	0.005(0.007)
0.40	0.006(0.006)	0.006(0.007)	0.006(0.007)	0.006(0.007)	0.006(0.007)	0.005(0.007)
0.45	0.005(0.006)	0.005(0.006)	0.005(0.006)	0.005(0.007)	0.006(0.007)	0.006(0.007)
0.50	0.006(0.006)	0.005(0.006)	0.005(0.007)	0.006(0.007)	0.006(0.007)	0.006(0.007)

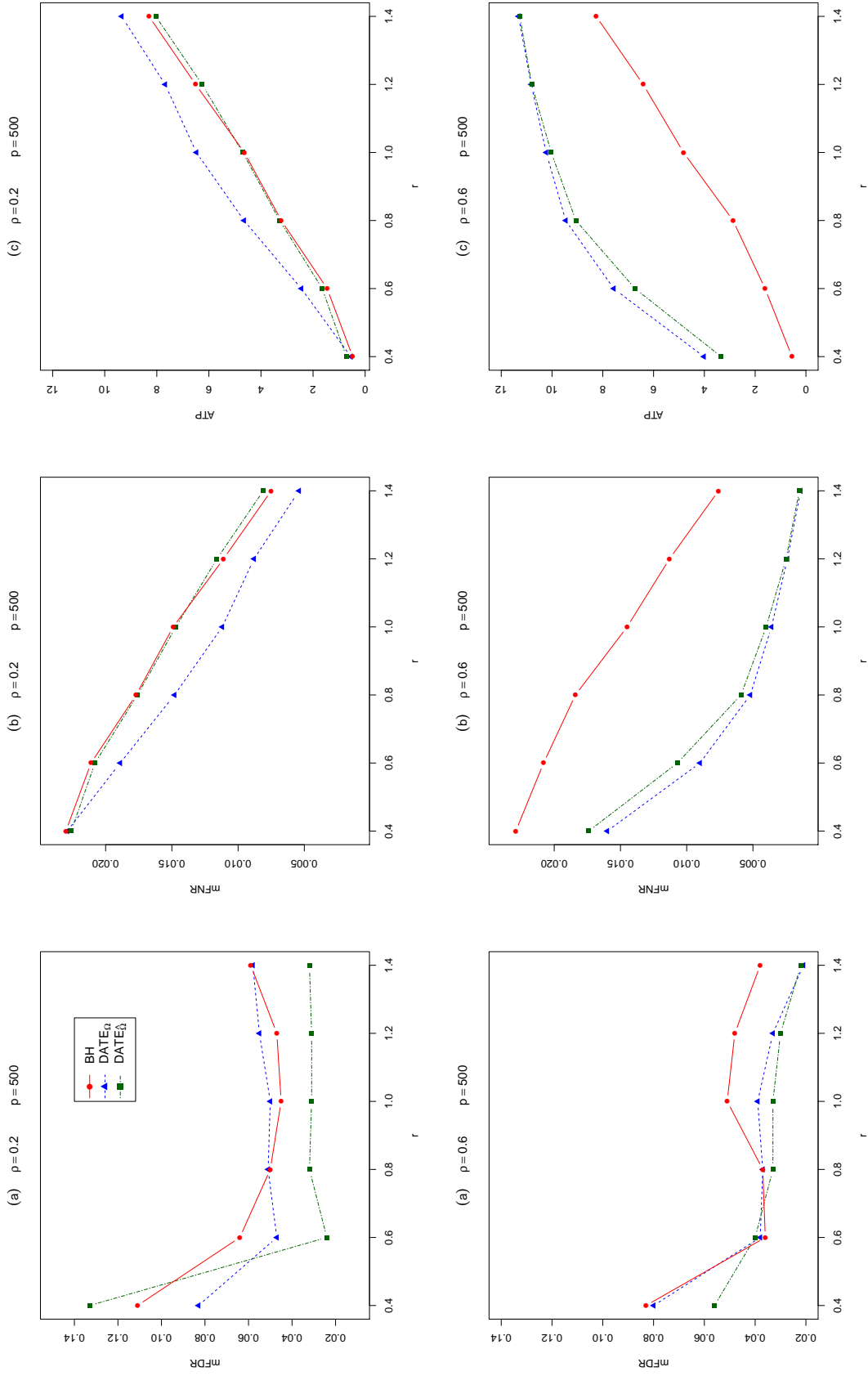


Figure 3: The mFDR, mFNR and ATP yielded by DATE<sub>Ω</sub>, DATE<sub>Ω̂</sub> and the BH procedure under model (a). The dimension  $p = 500$ , sample sizes  $n_1 = 60$  and  $n_2 = 60$  and  $\beta = 0.6$ .

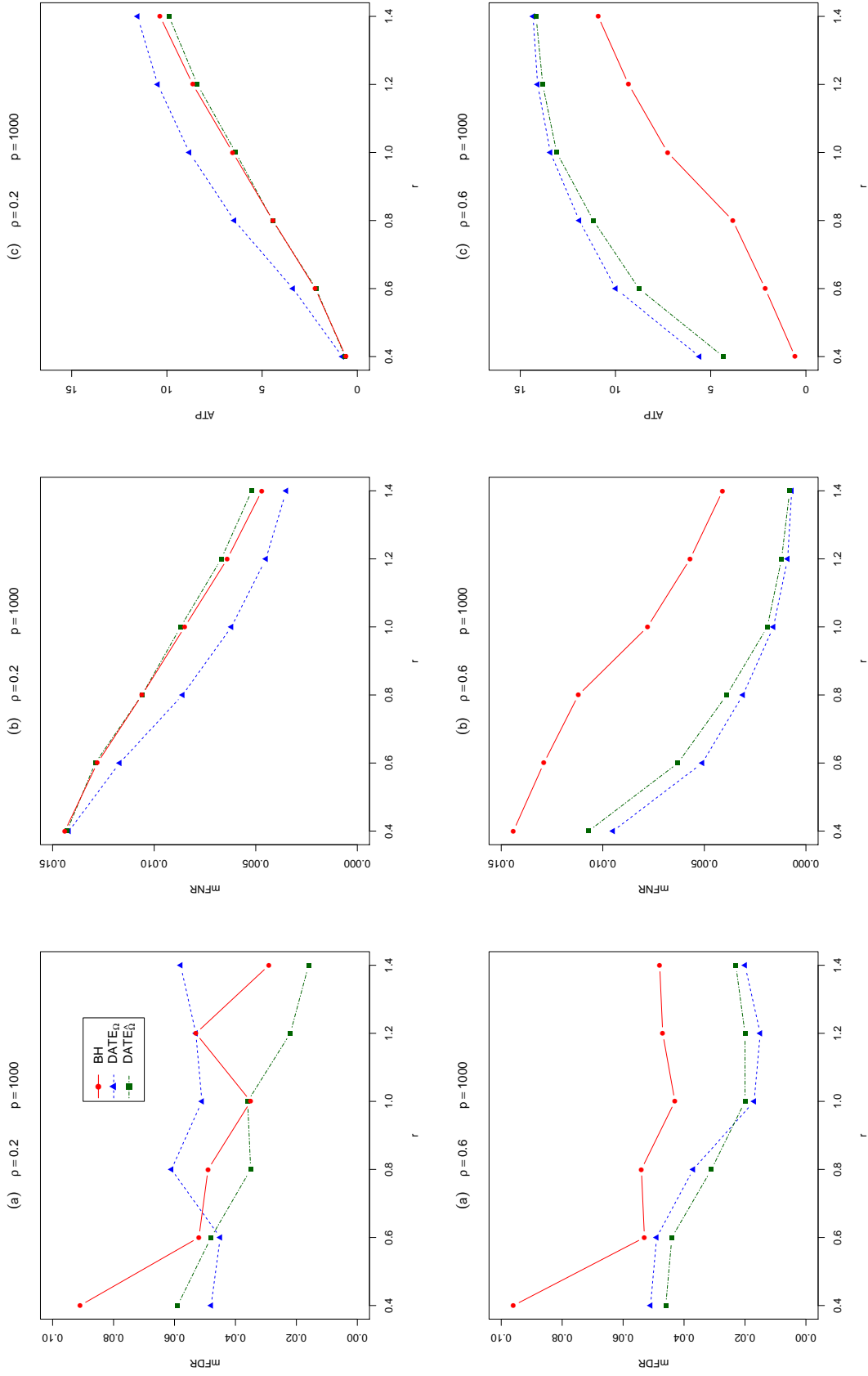


Figure 4: The mFDR, mFNR and ATP yielded by DATE<sub>0</sub>, DATE<sub>0</sub><sup>hat</sup> and the BH procedure under model (a). The dimension  $p = 1000$ , sample sizes  $n_1 = 60$  and  $n_2 = 60$  and  $\beta = 0.6$ .

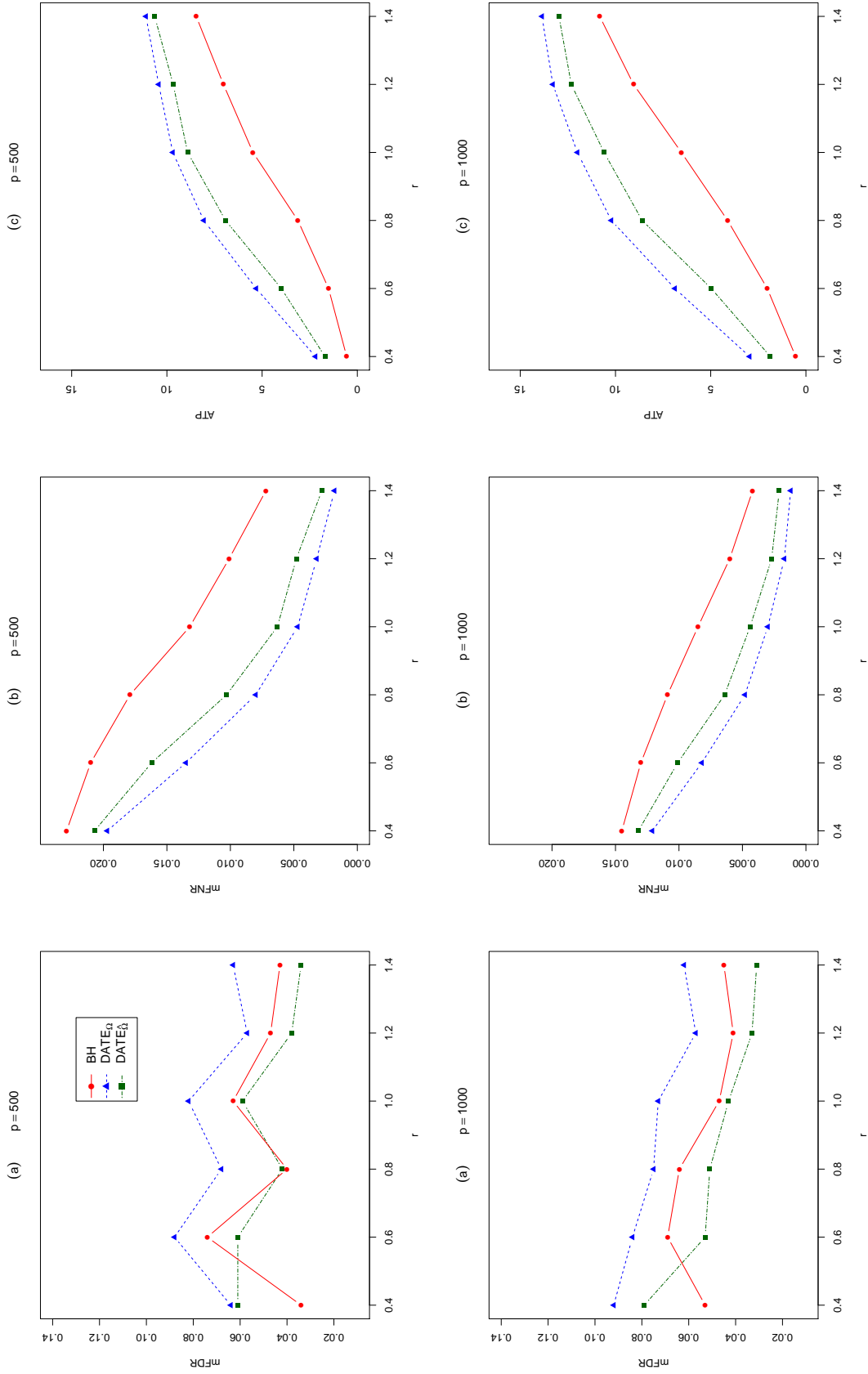


Figure 5: The mFDR, mFNR and ATP yielded by DATE<sub>Ω</sub>, DATE<sub>Ω̂</sub> and the BH procedure under model (b). The sample sizes  $n_1 = 60$  and  $n_2 = 60$  and  $\beta = 0.6$ .

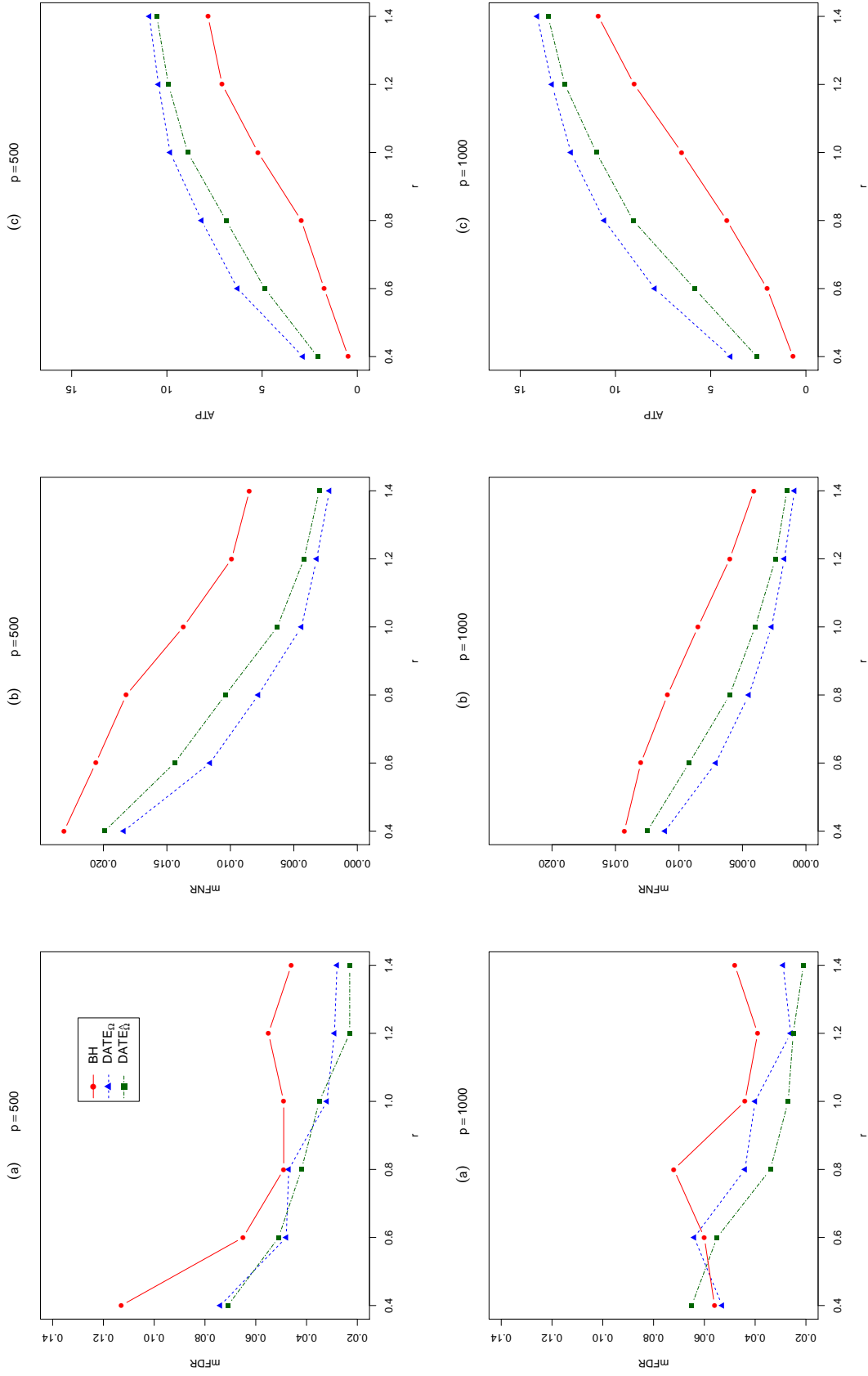


Figure 6: The mFDR, mFNR and ATP yielded by DATE<sub>Ω</sub>, DATE<sub>Ω̂</sub> and the BH procedure under model (c). The sample sizes  $n_1 = 60$  and  $n_2 = 60$  and  $\beta = 0.6$ .

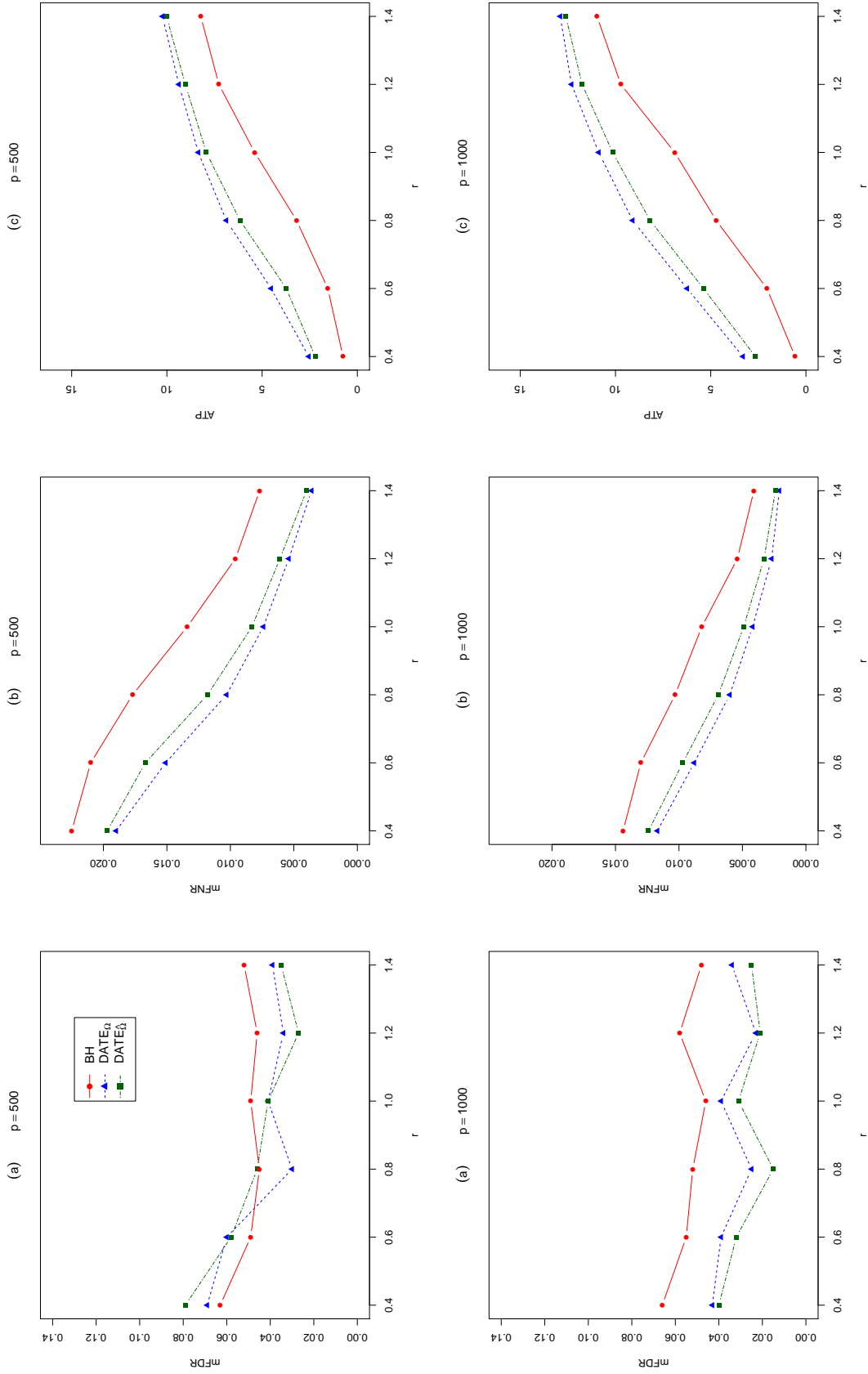


Figure 7: The mFDR, mFNR and ATP yielded by DATE<sub>Ω</sub>, DATE<sub>Ω̂</sub> and the BH procedure under model (d). The sample sizes  $n_1 = 100$  and  $n_2 = 100$  and  $\beta = 0.6$ .

Table 2: The number of differentially expressed genes identified by the BH, the DATE and both on chromosome X with the FDR controlled at the level of  $\alpha = 0.01, 0.005$  and 0.001.

FDR-controlled level	BH	DATE	Both
0.01	52	56	38
0.005	43	50	33
0.001	27	39	22

Table 3: The differentially expressed genes identified by the DATE not by the BH on chromosome X with the FDR controlled at level 0.001.

Gene symbol	Location	Description
PTCHD1	Xp22.11	patched domain containing 1
DMD	Xp21.2	dystrophin
SLC9A6	Xq26.3	solute carrier family 9 (sodium/hydrogen exchanger), member 6
KAL1	Xp22.32	Kallmann syndrome 1 sequence
TMSB15B	Xq22.2	thymosin-like 8
GPR64	Xp22.13	G protein-coupled receptor 64
ATP6AP1	Xq28	ATPase, H <sup>+</sup> transporting, lysosomal accessory protein 1
NXT2	Xq23	nuclear transport factor 2-like export factor 2
CLCN4	Xp22.3	chloride channel 4
VGLL1	Xq26.3	vestigial like 1 (Drosophila)
BEX1	Xq22	brain expressed, X-linked 1
SLC6A14	Xq23	solute carrier family 6 (amino acid transporter), member 14
BCOR	Xp21.2-p11.4	BCL6 corepressor
BCORL1	Xq25-q26.1	BCL6 corepressor-like 1
MUM1L1	Xq22.3	melanoma associated antigen (mutated) 1-like 1
SYTL5	Xp21.1	synaptotagmin-like 5
RLIM	Xq13-q21	ring finger protein, LIM domain interacting